La recherche d'information sur Internet au prisme de la théorie des facettes

Eric Boutin

Habilitation à Diriger des Recherches

La recherche d'information sur Internet au prisme de la théorie des facettes

Discipline : Sciences de l'Information et de la Communication

Laboratoire I3M

École doctorale de l'Université du Sud Toulon-Var

Présentée et soutenue publiquement le 9 Octobre 2008

Par Eric Boutin

- Jury -

- Patrick Abellard, Professeur
- Claude Baltz , Professeur
- Amos David, Professeur, Rapporteur
- Philippe Dumas Professeur, Tuteur de recherche
- Thierry Lafouge, Professeur, Rapporteur
- Sophie Pene, Professeur, Rapporteur
- Brigitte Trousse, Inria

Habilitation à Diriger des Recherches

La recherche d'information sur Internet au prisme de la théorie des facettes

Discipline : Sciences de l'Information et de la Communication

Laboratoire I3M

École doctorale de l'Université du Sud Toulon-Var

Présentée et soutenue publiquement le 9 Octobre 2008

Par Eric Boutin

- Jury -

- Patrick Abellard, Professeur
- Claude Baltz , Professeur
- Amos David, Professeur, Rapporteur
- Philippe Dumas Professeur, Tuteur de recherche
- Thierry Lafouge, Professeur, Rapporteur
- Sophie Pene, Professeur, Rapporteur
- Brigitte Trousse, Inria

Remerciements

Je tiens tout d'abord à remercier Philippe Dumas, mon Tuteur de Recherche. Il a su depuis dix sept ans m'accompagner, être là aux moments décisifs. Son optimisme naturel, son humanisme, sa sagesse, son esprit visionnaire, son sens du consensus, m'ont permis d'évoluer dans un environnement stable et m'ont donné des repères précieux.

Mes remerciements vont également aux rapporteurs qui, chacun avec leur sensibilité m'ont permis d'enrichir ce travail et d'entrevoir des perspectives nouvelles.

Je remercie également les membres du jury qui me font l'honneur de lire le mémoire et d'assister à cette soutenance.

J'ai une pensée reconnaissante pour les doctorants, collègues ou parents qui ont eu la gentillesse bienveillante de consacrer du temps à cette tâche ingrate qu'est la relecture du manuscrit.

Je remercie chaleureusement ma femme, Laurence et mes deux enfants Manon et Yannis pour avoir créé les conditions favorables à l'épanouissement de cette recherche.

Je remercie les doctorants que j'ai co-encadrés (Pédro, Roberto, Eve, Franck, Philippe, Jean Dominique, Jean Pierre) et ceux que je co-encadre aujourd'hui (Natacha, Guillaume, Pei, Mohamed, Hajer, Stéphane, Maher). Ils m'ont permis de gagner en maturité dans la discipline.

J'ai une pensée aussi pour les collègues chercheurs avec lesquels l'échange scientifique est source de bouillonnement intellectuel toujours stimulant.

Je vous remercie enfin, vous lecteur qui découvrez ce texte. Puisse sa lecture renforcer les passerelles entre univers étanches et contribuer à la fertilisation croisée des disciplines.

SOMMAIRE

PREAMBULE

Les occasions sont rares pour un chercheur de porter un regard réflexif sur sa production scientifique. L'objet de ce préambule est précisément de poser les valises et de regarder le chemin parcouru. Nous proposons d'analyser ce chemin à travers trois prismes :

- Positionnement et évolution de la recherche en Sciences de l'Information et de la Communication (SIC)
- Cohérence du parcours et dynamique de recherche
- Collaborations scientifiques suscitées par cette recherche

Chacun de ces prismes offre une grille de lecture possible et permet d'éclairer le présent document.

A. Positionnement et évolution de la Recherche en SIC

Le premier prisme est celui de l'évolution de cette recherche et de son inscription au sein des travaux en Sciences de l'Information et de la Communication. On s'est intéressé (Dumas et al. – 2005, Gallezot et al. – 2006) à la représentation des travaux en SIC le long d'un continuum multidimensionnel entre information et communication. Cette vision bipolaire est réductrice mais peut servir d'éclairage et permettre de mieux comprendre l'évolution de ma recherche depuis 13 ans et son positionnement actuel.

En 1995, une première communication[1] (Boutin et al. – 1995a) effectuée au colloque de l'ISSI[2] correspond à un travail qui s'inscrit en science de l'information. Treize ans plus tard, on constate une évolution vers une prise en compte d'une dimension communication plus marquée. Cette évolution

[1] dont le titre est : "a new approach to display real co-authorship and co-topicship through network mapping"

[2] International Conference of the International Society for Scientometrics and Informetrics

correspond à une opportunité scientifique pour un chercheur des SIC en France s'il veut exister dans la communauté internationale.

1. Positionnement international de la communauté des SIC

La communauté des SIC en France se singularise par le regroupement, au sein d'une même discipline, de recherches en information et communication là où, à l'étranger, ces disciplines sont souvent séparées (on pense notamment à la distinction anglo-saxonne entre *information sciences*, *communication sciences* et *media studies*).

Ce regroupement français est considéré par les auteurs en SIC soit comme une force, soit comme une faiblesse selon le point de vue qu'ils adoptent. Nous pensons que cette particularité française est une faiblesse qui peut être transformée en opportunité.

- Faiblesse car à l'heure de la compartimentation des savoirs, les laboratoires étrangers se concentrent sur quelques thématiques et y associent les moyens humains et financiers. Ils atteignent ainsi une taille critique. Difficile pour un chercheur français de rester dans le « front de recherche »[3] sur son sujet quand les crédits et les hommes sont éparpillés sur des thématiques beaucoup plus élargies.
- Opportunité pour autant qu'on puisse avoir un état d'esprit et une démarche de décloisonnement.

2. Fragmentation des savoirs

Pierret (2006) reprend la logique de Swanson (1993) qui décrit la communauté scientifique comme s'organisant en créant des disciplines toujours plus pointues. « Les scientifiques, qui ont depuis longtemps abandonné l'idée de pouvoir lire tous les écrits qui paraissent, se sont organisés pour travailler en spécialités, permettant à chacun de se concentrer sur une petite partie de la

[3] Le concept de front de recherche a été introduit par De Solla Price (1965) pour désigner un ensemble de documents émergents d'un champ de recherche. Pour Chen (2006), le front de recherche est l'état de l'art d'une spécialité à un instant t. Persson (1996) distingue le front de recherche constitué des articles citants et la base intellectuelle (cœur de la littérature scientifique) composée des articles cités.

littérature. Les spécialités qui croissent trop vite se divisent à leur tour en sous spécialités. Ainsi le volume de chaque spécialité est à peu près constant laissant à chacun l'illusion de suivre la littérature de son domaine ». Le graphe de la *Figure 1* (réalisé d'après Swanson) rend compte de la fragmentation des connaissances au fur et à mesure qu'elles s'accroissent.

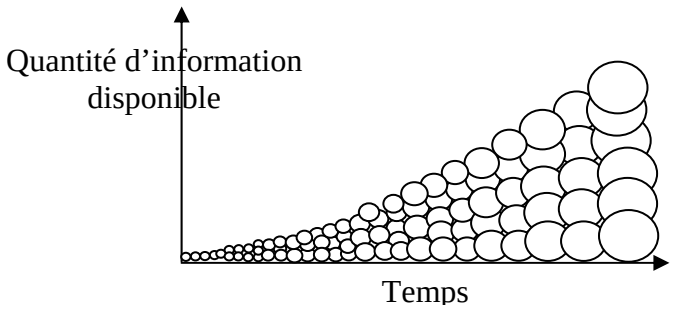


Figure 1 : Evolution du volume d'information disponible au cours du temps

Pour Morin (1999), « le principe de séparabilité s'est imposé dans le domaine scientifique par la spécialisation, puis il s'est dégradé en hyperspécialisation et compartimentation ». En divisant ainsi la production scientifique, les relations logiques entre différentes spécialités tendent à être négligées, ignorées ou occultées. Or, deux spécialités peuvent sur certains points être complémentaires et porteuses, ensemble, de nouveaux savoirs. La fragmentation du savoir cache ces relations pertinentes.

3. Passerelles entre information et communication

Le travail de Don Swanson porte sur l'identification de ces relations à travers l'exploitation des bases de données bibliographiques. Il cherche à identifier l'*undiscovered public knowledge* [Swanson, 1986].

Son travail montre qu'il y a un potentiel à transférer, dans certains univers, des idées développées dans d'autres. Plus on assiste à une compartimentation des savoirs, plus il est important de penser des logiques transversales de passerelles entre ces savoirs compartimentés. De ce point de vue, la spécificité française des SIC offre au chercheur un gisement insoupçonné de transfert entre champs de recherche. Alors que les communautés sont disjointes dans d'autres pays, elles sont regroupées en France au sein des SIC. On retrouve cette diversité

dans un laboratoire en SIC comme I3M où cohabitent des chercheurs qui ont des centres d'intérêts très différents. Cette cohabitation se traduit par des rencontres, des discussions scientifiques qui peuvent conduire à des collaborations scientifiques. Il est plus facile pour un chercheur français en sciences de l'information de transposer à son domaine une idée empruntée aux sciences de la communication qu'à un chercheur américain en *information science* d'emprunter un concept en *communication science*. Telle est sans doute la clé d'un positionnement de la recherche française en SIC sur la scène internationale.

4. Nécessité d'un brassage scientifique

Une illustration peut être fournie. J'ai été associé au jury de thèse de David Reymond (2007) qui concerne l'analyse webométrique[4] de sites web universitaires. Difficile de positionner une recherche sur la scène internationale sur cette question quand on découvre la littérature très abondante de certains chercheurs anglo-saxons sur le sujet (Thelwall par exemple). La singularité du travail proposé est de se positionner non plus sur une logique webométrique traditionnelle mais sur une approche qui emprunte à la webométrie tout en se nourrissant par exemple de l'approche sémiotique. Il y a là une piste de recherche originale qui correspond à une stratégie de niche.

5. Un métissage scientifique revendiqué

Ma recherche doctorale et les recherches que j'ai initiées ou encadrées sont souvent nourries de pistes transposées d'autres domaines de la connaissance :

L'analyse réseau, initialement développée dans le domaine de la sociologie est mise au service de la bibliométrie et de l'intelligence économique dans ma thèse (Boutin 1999).

[4] Björneborn et Ingwersen (2004) définissent la webométrie, comme « L'étude des aspects quantitatifs de la construction et de l'utilisation de ressources informationnelles, de la structure de ces informations et des technologies du web basées sur des approches infométriques et bibliométriques ».

La recherche de Liu Pei, dont je co-encadre la thèse, (Boutin et al. – 2007a, Liu et Al. – 2008, Boutin et al. – 2008a) est un regard sur la culture asiatique de l'intelligence économique.

Le travail de doctorat de Natacha Romma dont je co-encadre les travaux correspond aussi à cet esprit. Avec la montée en puissance des outils communicationnels sur Internet, elle s'intéresse à la transposition du paradigme de la communication engageante[5] (Bernard et Joule, 2004 ; Bernard et Joule, 2005) aux nouveaux dispositifs numériques.

La thèse de Roberto Penteado, soutenue sous mon co-encadrement (2006) est un trait d'union entre information et communication : le chercheur, journaliste de formation, revisite l'analyse bibliométrique pour l'appliquer aux médias.

Ce travail d'habilitation est également un travail qui va puiser ailleurs et qui rapporte aux SIC. Il transpose une théorie développée dans les années 1930 par Ranganathan à l'ère numérique; il mobilise des approches issues de travaux en linguistique computationnelle, en psychologie cognitive, et du domaine de la « recherche d'information »[6].

Pour conclure sur ce point, nous revendiquons une posture de métis scientifique. C'est une position scientifique difficile à tenir. Elle consiste à nourrir les sciences de l'information d'apports appartenant à des disciplines connexes. Le problème est qu'il est impossible d'avoir dans ces disciplines connexes le niveau de finesse que l'on peut avoir dans sa discipline d'origine. Le métis scientifique est entre deux. Il est parfois rejeté du côté où il emprunte car son implication récente dans le champ le conduit souvent à un manque de rigueur qui lui est reproché par ceux qui consacrent leur recherche à ce sujet. Il est souvent isolé dans son propre champ de par l'originalité de l'approche qu'il

5 Ce paradigme prend appui sur la théorie de l'engagement (Joule et Beauvois, 1987 - 1998) qui montre comment l'engagement dans des situations de communication comportant des pratiques de médiation et de médiatisation permet d'inscrire des acteurs (individus et collectifs) dans des cours d'action.

6 Nous mettons l'expression « recherche d'information » entre guillemets car nous l'employons dans ce mémoire non pas dans son sens courant mais comme traduction littérale du terme anglosaxon d'information retrieval.

présente. La compartimentation des savoirs est à cet égard beaucoup plus rassurante.
L'existence des SIC comme discipline scientifique est une condition nécessaire mais non suffisante de ce métissage. Le brassage entre information et communication doit se faire au sein de la section et au sein des laboratoires pour arriver à cette fertilisation croisée. Nous considérons que le métissage scientifique est une stratégie de niche qui est à même de donner une visibilité des travaux de notre communauté à l'international.

B. Cohérence du parcours et dynamique de recherche

Entre les premiers travaux publiés en 1995 sur l'application de l'analyse réseau dans des contextes bibliométriques[7] et les travaux les plus récents, il y a une prise de distance par rapport au sujet de thèse. Notre objet est ici de montrer la cohérence de ce parcours de recherche. L'évolution s'est opérée par glissements successifs avec des points d'articulation permettant de passer d'une étape à l'autre. Nous aimerions faire ressortir les jalons de cette évolution autour de quatre grandes périodes chronologiques.

1. Analyse réseau : bibliométrie au service de l'Intelligence Economique[8] depuis 95

Ce parcours de recherche s'est construit au départ autour d'un projet de thèse développé au sein du laboratoire le Pont-Crrm (Aix Marseille 3) et dirigé par Philippe Dumas et Hervé Rostaing. Cet ancrage de départ au niveau d'un laboratoire très centré autour de problématiques opérationnelles de « l'intelligence économique », de la veille stratégique, du développement d'outils et de méthodes au service de l'info-décision va se retrouver dans une

[7] Quoniam (1996) définit l'analyse bibliométrique comme l'application, à des données textuelles issues de bases de données bibliographiques, de l'analyse automatique de données textuelles par des méthodes statistiques et d'analyse pour la prise de décision.
[8] D'après Alain Juillet (2006), "L'intelligence économique consiste en la maîtrise et la protection de l'information stratégique pour tout acteur économique. Elle a pour triple finalité la compétitivité du tissu industriel, la sécurité de l'économie et des entreprises et le renforcement de l'influence de notre pays."

recherche action qui a pour objectif de fournir, à travers « l'analyse réseau », des indicateurs permettant aux décideurs d'appréhender les données massives qu'ils ont à traiter. L'analyse réseau a été développée en sociologie des organisations dans la logique des représentations de type sociogramme de Moreno (1954). Nous avons proposé dans la thèse de présenter l'analyse réseau comme outil infométrique permettant de représenter une information complexe par différentes cartographies et indicateurs offrant chacun une grille de lecture des documents primaires qu'ils ont à décrire.

Les publications scientifiques de cette période (Quoniam & al. – 1995, Boutin & al. – 1995a, 1995b, 1996a, Panteado & al. – 2003a, 2003b, 2005, Dumas & al. – 2005, Gallezot & al. – 2006, Panteado & Boutin – 2007, Panteado & al. – 2007a, 2007b) sont effectuées en association avec des chercheurs confirmés.

2. L'analyse réseau comme outil d'analyse de données : 1996-2004

Une fois l'analyse réseau utilisée dans le contexte de l'analyse bibliométrique, l'idée est alors de s'abstraire de ce contexte applicatif pour généraliser l'approche à des contextes caractérisés par le traitement automatique de corpus volumineux de données. Des coopérations avec des chercheurs en sciences de gestion permettent de montrer l'intérêt de l'approche en analyse de données. L'analyse réseau est alors intégrée dans la boite à outil du statisticien (Boutin & al. – 1996b, 1997a, 1997b, 2000, Rostaing & al. – 1998, Boutin & Ferrandi – 1999a, 1999b, Bertacchini & al. – 2000, Ferrandi & Boutin – 2001, Gasté & al. – 2003b, Boutin & Gasté – 2004).

3. De la bibliométrie relationnelle à la webométrie relationnelle : 1998-2008

Nous avons souhaité appliquer nos techniques bibliométriques à la webométrie relationnelle en analysant les interactions hypertextuelles entre sites web.

Ma recherche en ce domaine reste profondément marquée par une participation active à la création et au développement d'une startup Internet. Cette expérience industrielle a mobilisé beaucoup de mon énergie de 2000 à 2003 en parallèle de mes charges d'enseignement et de recherche. Soutenue par le

Ministère de la Recherche et lauréate en 2000 et 2001[9] du concours de création d'entreprise innovante, elle poursuit l'objectif ambitieux de développer un indicateur de pertinence de moteur de recherche web qui soit meilleur que celui de Google. Au niveau technologique, le défi est relevé en moins de deux ans. Cette technologie se présente sous la forme d'un booster de pertinence pour moteurs de recherche conventionnels leur permettant d'arriver à rattraper la pertinence de Google. Elle est donc susceptible d'intéresser des moteurs de recherche compétiteurs de Google (En 2002 Altavista, Yahoo, Fast, Inktomi essentiellement). Au niveau économique, l'expérience bute sur une certaine frilosité des investisseurs français et sur une évolution du secteur des moteurs de recherche dans le sens d'une concentration forte des acteurs. A l'heure où la startup démarche ses clients, les concurrents majeurs de Google sont en pleine restructuration de leur actifs : Yahoo, Altavista, Fast AlltheWeb, se regroupent pour ne constituer qu'une seule entité et l'heure est plus à la rationalisation des équipes et à la création de synergies entre des équipes, jusqu'à présent concurrentes, qu'à l'absorption d'une nouvelle technologie.

Cette position de responsable de recherche de cette entité m'a donné l'opportunité unique d'être au cœur de la recherche-appliquée dans le domaine des indicateurs de pertinence de moteur de recherche. Cette expérience, très enrichissante, laisse des traces durables dans les savoir faire acquis tout autant qu'elle crée des pistes de recherches nombreuses qui font l'objet de développements actuels. Cette expérience va confirmer une recherche action qui, tout en s'inscrivant dans la continuité de la thèse, se focalise désormais sur une vision web plus prononcée.

Le monde du web a été très rapidement choisi comme terrain d'expérimentation privilégié pour deux autres raisons : Les données web représentent un défi par rapport aux données disponibles dans les bases de données bibliographiques. En effet alors que les données des bases de données bibliographiques sont structurées, les données du web sont peu ou mal structurées. D'autre part, le web correspond à une source d'information ouverte de plus en plus utilisée dans des contextes de recherche d'information grand

[9] http://www.adminet.com/jo/20011120/RECT0100276A.html

public ou professionnels. Il devient alors intéressant de passer de la bibliométrie à la webométrie et de transposer au web des logiques d'analyses relationnelles automatiques mal prises en compte par les outils de recherche.

Le passage de l'analyse réseau comme outil d'analyse bibliométrique à l'analyse réseau comme outil d'analyse webométrique conduit aussi à un changement profond d'objet d'étude. Typiquement en bibliométrie, les objets étudiés sont des auteurs dont on étudie la co-signature, des mots clés dont on recherche la co-occurrence ou des documents dont on va mesurer la proximité.

L'analyse réseau appliquée au domaine webométrique débouche sur des représentations cartographiques représentant l'interaction hypertextuelle entre les sites web d'un corpus. La question est de savoir ce qu'on peut inférer de ces représentations hypertextuelles sur la logique des relations réelles entre les objets. Pour répondre à cette question, il faut s'interroger sur la ou les significations du lien hypertexte. Ce travail fera l'objet de nombreuses validations expérimentales et de contrats de recherche industriels[10]. Boutin & al. – 1998, 1999c, Rostaing & Boutin – 1999, Bertacchini & Boutin 2003, Boutin & Martaillan – 2005, Boutin & al. – 2008d,

Durant cette période, je co-encadre divers doctorants qui vont consacrer leurs thèses à approfondir l'analyse réseau dans le domaine du web. On peut citer le cas de l'intelligence territoriale avec le travail de Guillaume Perrin (Boutin & Perrin – 2005, Perrin & Boutin – 2005, 2006) ou de l'Intelligence informationnelle avec Natacha Romma (Romma & Boutin – 2006), l'analyse du lien hypertexte avec Mohamed Youssef (Youssef & al. – 2008a, 2008b).

4. De la webométrie relationnelle à la webométrie centrée utilisateur : 2004-2008

Cette étape se caractérise par un changement de vision assez radical. Dans les trois premières étapes, le cœur est la technique d'analyse relationnelle qu'il s'agit de mettre au point puis de déployer sur divers corpus de données

[10] notamment une étude dont j'ai été le porteur sur la mesure de l'internet public en région PACA. Commanditée par la Préfecture de Région et le Conseil Régional, cette étude consistait à recenser les acteurs et institutions publiques sur le web qui dépendent de la région PACA.

structurées puis non structurées, bibliographiques puis webométriques dans des contextes partant de l'intelligence économique pour aller vers l'analyse de données. Désormais, nous adoptons une vision caractérisée en ce qu'elle est désormais centrée sur l'utilisateur. L'utilisateur qui nous intéresse est toute personne qui recherche de l'information sur Internet. Nous privilégions plus précisément la personne qui recherche l'information sur Internet dans un contexte professionnel, dans une logique d'intelligence économique. L'objectif est alors d'identifier correctement le besoin de l'utilisateur et de chercher dans la boite à outils l'analyse la mieux adaptée pour répondre au problème posé : l'analyse relationnelle en est une, l'analyse textuelle une autre. L'objectif est alors d'être capable de mobiliser la bonne analyse de façon judicieuse. (Dumas & al. – 2002, Pierret & Boutin – 2004, Pierret & al. – 2005, Boutin & Cadel – 2005, Cadel & Boutin – 2005, Romma & Boutin – 2005a, 2005b, Boutin & Romma – 2005, Boutin & al. – 2006a, 2006b, Boutin & Quoniam – 2006, Perrin & al. – 2006, Boutin & al. – 2007a, Liu & al. – 2008, Boutin & al. – 2008a, 2008b, 2008c, Ertzscheid & al. – 2008).

Le Tableau 1 illustre de façon schématique les 4 étapes de cette recherche et l'évolution des publications par thème et par année. Une petite case grisée correspond à une publication pour l'année et le thème considéré.

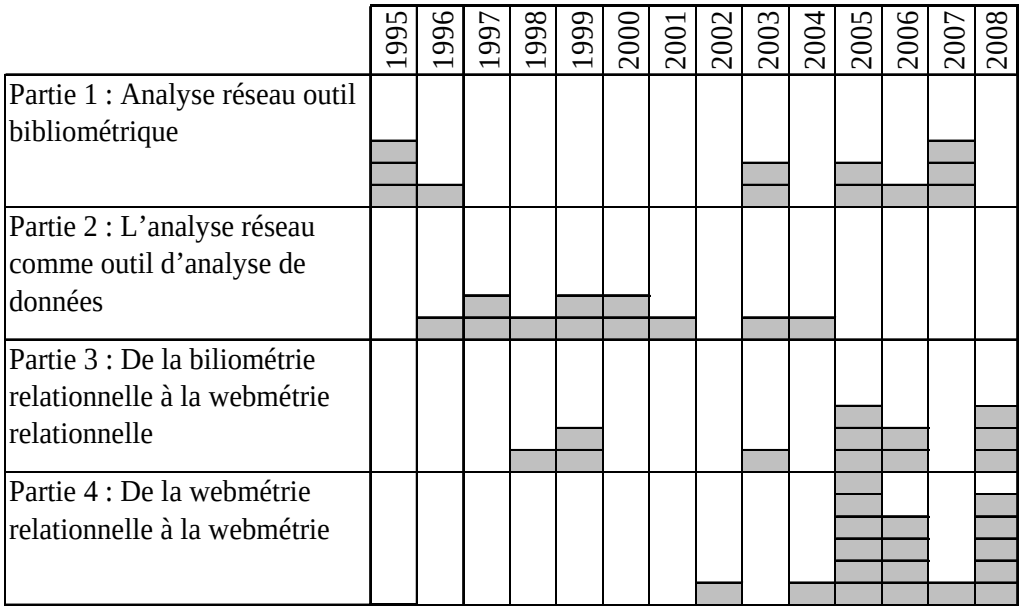


Tableau 1 : Les grandes étapes de la recherche : approche chronologique

Pour résumer ces 4 approches temporelles, on peut dire que chacune d'elle privilégie une triade particulière correspondant à une source d'information, un outil d'analyse et un domaine applicatif. Le Tableau 2 comporte des flèches qui illustrent le basculement d'une période à l'autre.

Etape	Domaine applicatif	Source d'information	Outil d'analyse
1- Analyse réseau outil bibliométrique au service de l'IE	Bibliométrie, scientométrie, intelligence économique	Bibliographique	Analyse réseau
2- L'analyse réseau comme outil d'analyse de données	Analyse de données Statistique marketing	Bibliographique Fichiers logs Enquêtes marketing	Analyse réseau
3- De la biliométrie relationnelle à la webmétrie relationnelle	Intelligence économique et territoriale	Corpus de pages web Forums de discussion	Analyse réseau
4- De la webmétrie relationnelle à la webmétrie.	Intelligence économique	Corpus de page web	Indicateurs spatiotemporels et recherche de signaux faibles

Tableau 2 : Glissement de la recherche par étape

On observe dans ce parcours trois glissements successifs. Le premier permet de passer de l'utilisation de l'analyse réseau en bibliométrie à son utilisation en analyse de données. Le deuxième glissement permet de passer d'une source d'information bibliographique à une source d'information web. Le dernier glissement traduit le passage d'une dimension outil à une dimension analyse.

Le sujet de la présente habilitation traduit une approche qui concerne la recherche d'information sur internet d'un point de vue centré sur l'utilisateur et ne mobilisant plus forcément exclusivement l'analyse réseau. Elle s'inscrit en cohérence avec la logique qui sous-tend ce parcours de recherche.

C. Prisme des collaborations scientifiques associées à cette recherche

Cette recherche ne peut pas être dissociée du contexte humain qui l'a nourrie. La co-publication d'articles avec des acteurs de la communauté a toujours été privilégiée. Ayant travaillé sur l'analyse réseau, j'étais bien placé pour mettre en pratique le principe de collaboration dans mes travaux de recherche ; j'y ai été poussé dès le départ par les chercheurs de l'équipe du Crrm. Au départ, ces co-publications avaient lieu avec mon directeur, co-encadreur et membres du laboratoire. Depuis une période plus récente, mes co-auteurs sont aussi des doctorants ou des chercheurs dans et hors champ des SIC en France et à l'étranger. Au total cette recherche se traduit en mai 2008 par 61 publications scientifiques avec 45 chercheurs différents.

La Figure 2 permet de visualiser les auteurs avec lesquels j'ai collaboré. Il y a 45 sommets sur ce graphe correspondant aux 45 personnes avec lesquelles j'ai publié au moins un article. J'ai symbolisé à gauche du réseau un sommet fictif rayonnant qui serait relié aux 45 autres. Nous n'avons pas mentionné ce sommet pour faciliter la lisibilité du graphe. Les formes des sommets correspondent au statut des auteurs de ces publications. Un carré correspond à un enseignant chercheur, un rond à un doctorant, un triangle à un professionnel. La couleur rouge permet d'identifier les 11 acteurs avec lesquels j'ai collaboré et dont j'ai co-encadré les travaux. Les deux auteurs représentés dans des carrés rouges sont donc des anciens doctorants dont j'ai co-encadré les travaux et qui ont maintenant un statut d'enseignants chercheurs. Un lien entre deux sommets signifie que j'ai publié au moins un article conjointement avec ces deux auteurs.

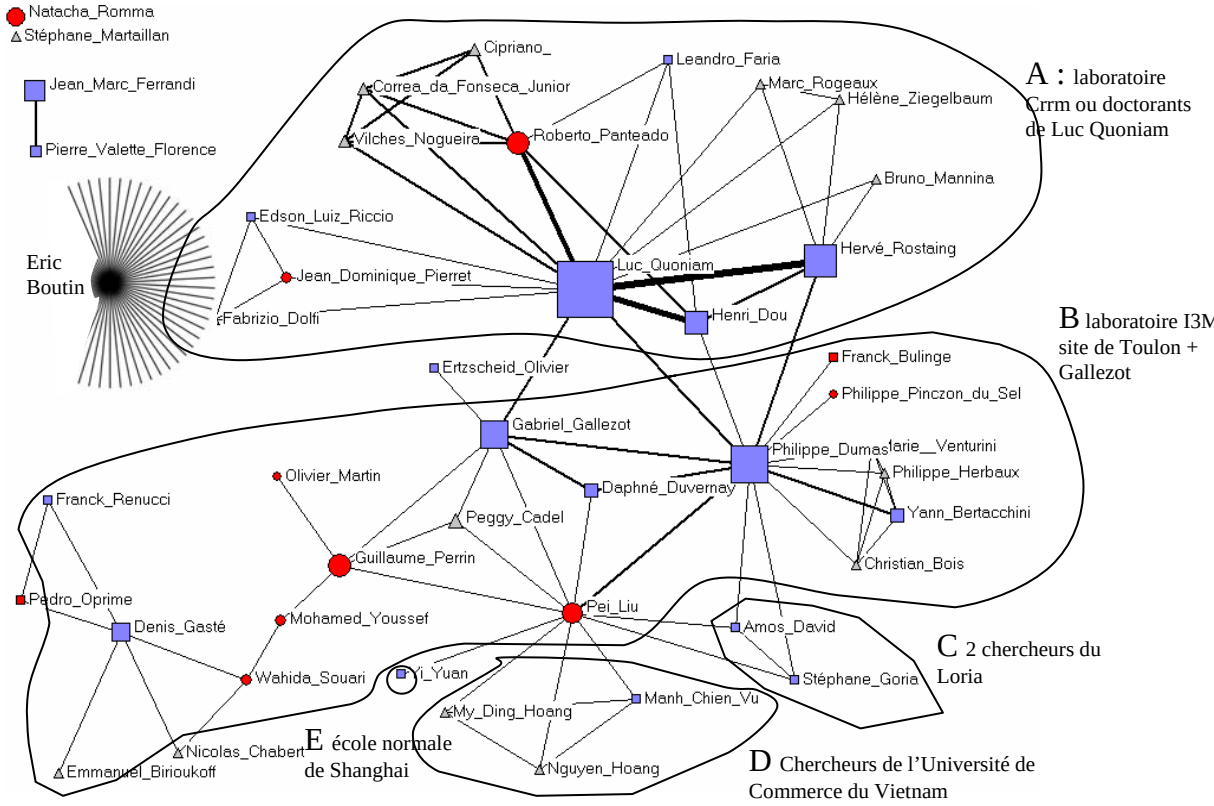


Figure 2 : réseau des collaborations d'auteurs sur la période 1995-2008

Ce graphe présente 4 composantes fortement connexes. Au nord-ouest du graphe, on observe deux chercheurs isolés : cela correspond à des recherches cosignées par uniquement deux auteurs. L'association Jean-Marc Ferrandi Pierre-Valette Florence correspond à deux chercheurs en sciences de gestion. Le reste du graphe se présente sous forme d'une composante fortement connexe de 41 auteurs. On observe au Nord un groupe (noté A) correspondant à des personnes ayant appartenu au Crrm ou fait une thèse dirigée par Luc Quoniam. Ces collaborations correspondent à des travaux dans le domaine de l'analyse de données textuelles. Le groupe du milieu (noté B) correspond à des auteurs appartenant au laboratoire I3M[11]. Mis à part Gabriel Gallezot, tous ces auteurs appartiennent au site de Toulon. Le groupe noté C correspond à une collaboration avec des chercheurs de l'équipe du Loria. Le groupe noté D correspond à une collaboration avec des chercheurs de l'Université de

[11] Le laboratoire I3M, "Information, Milieux, Médias, Médiations", Equipe d'Accueil 3820, est né en 2004. Situé sur les deux sites de l'Université de Nice-Sophia Antipolis et de l'Université du Sud Toulon-Var, il est co-dirigé par les Professeurs Philippe Dumas (Toulon) et Paul Rasse (Nice).

Commerce du Vietnam. Le point noté E correspond à une collaboration scientifique avec un chercheur de l'école normale supérieure de Shanghai.

Nous avons dans un second temps souhaité visualiser Figure 3 les auteurs avec lesquels trois publications scientifiques au moins ont été écrites. Ce réseau comporte 14 acteurs appartenant au laboratoire primitif d'accueil d'Eric Boutin (Crrm) et à son laboratoire actuel (I3M). La valeur figurant sur chaque arc correspond à l'intensité de la relation c'est-à-dire au nombre de publications écrites conjointement par les deux auteurs (situés aux extrémités de l'arc).

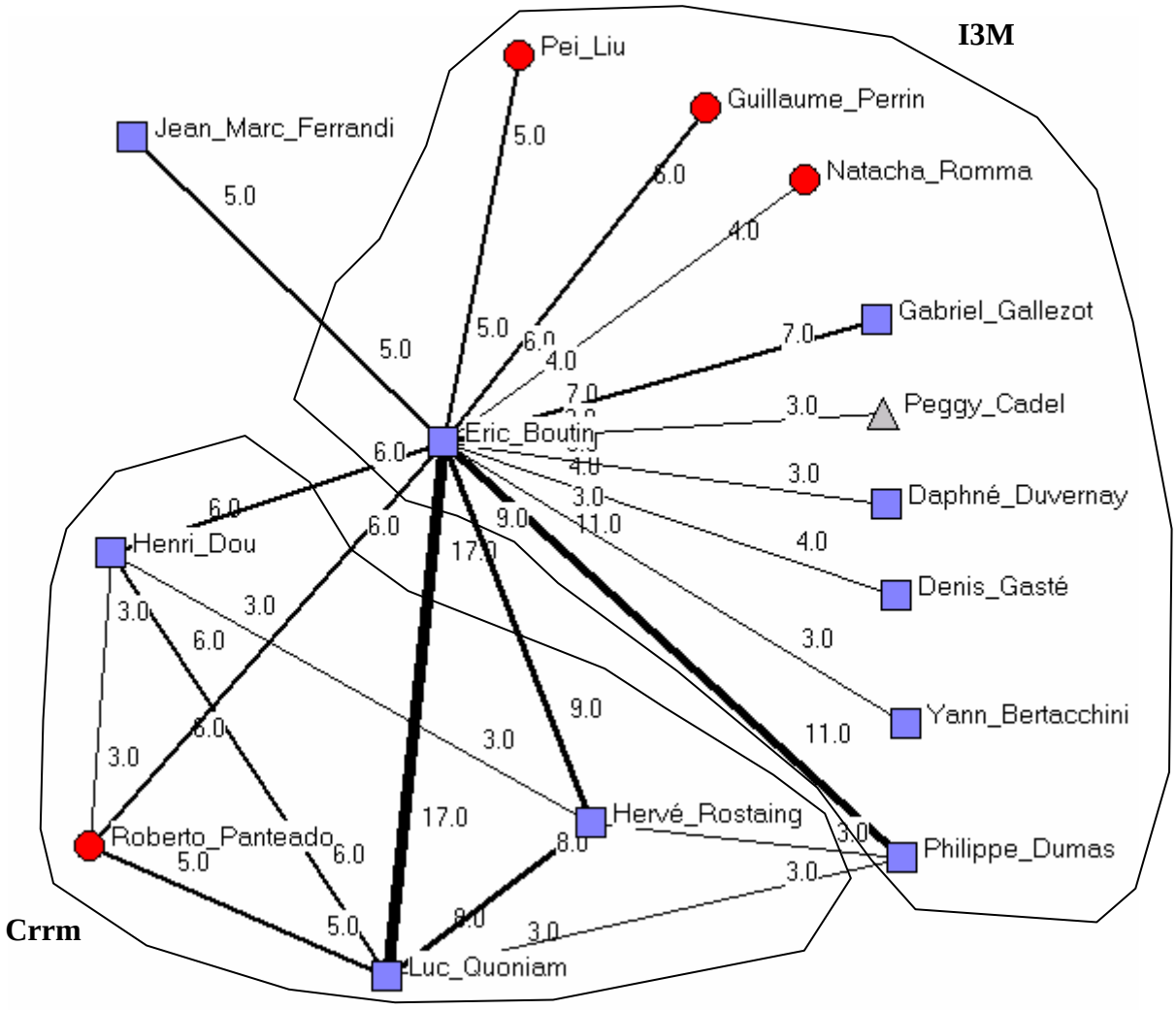


Figure 3 : visualisation du réseau des auteurs ayant écrit au moins trois publications avec Eric Boutin

INTRODUCTION :

« Chaque sujet a une capacité finie de lecture claire. C'est déjà bien beau, on a une petite portion de lecture claire et distincte, le reste on bafouille. On enveloppe le monde entier, oui, mais confusément, obscurément, d'une manière illisible. Et on a notre petite portion, notre petite lueur claire et distincte, notre petite lueur sur le monde, notre petite région de monde: ma chambre à moi. C'est déjà pas mal si j'enveloppe ma chambre à moi! Il ne faut pas demander beaucoup plus. »

Cours de Gilles Deleuze
16/12/1986
Leibniz Le pli Récapitulation

A. La gestion de l'univers informationnel : regards croisés

La question de la gestion de l'univers informationnel peut être étudiée à travers le prisme de travaux multidisciplinaires dans les domaines par exemple de la psychologie cognitive[12], de l'analyse de données, de la philosophie, de l'art, de l'économie, des sciences de gestion, des sciences de l'information et de la communication, de l'informatique. Ces travaux tant théoriques que pratiques privilégient, selon le cas, le regard sur des dispositifs techniques ou les dispositifs socio-techniques[13] à mettre en place comme lieu de rencontre de l'offre et de la demande d'information, le regard sur les usages (par exemple

[12] Le point de vue de la psychologie cognitive s'intéresse par exemple à la surcharge infomationnelle de l'usager et aux stratégies de biais cognitifs qu'il met en œuvre pour les dépasser.

[13] La notion de Dispositif Socio-Technique d'information Communication est considérée comme un «lieu privilégié d'interaction entre communication et transmission. Il est d'une part bien plus qu'un moyen permettant la transmission (ce terme étant utilisé dans son acception temporelle) ou la communication (ce terme étant utilisé dans son acception spatiale). Chaque Distic, par exemple Internet, se caractérise par une articulation spécifique et nouvelle entre transmission et communication. Il est d'autre part bien plus qu'un ensemble structuré de moyens langagiers pluri-sémiotiques mis en œuvre par une instance de production pour réaliser des intentions communicatives. En envisageant la communication comme des interactions production-dispositif-réception se réalisant au sein de multiples contextes, le dispositif est conçu comme un lieu de médiation, composé de multiples facteurs sémiotiques, esthétiques et techniques en interaction qui relient sensoriellement et de manière médiate les acteurs sociaux ». (Source : site du laboratoire I3M: http://i3m.univ-tln.fr/Une-definition-du-Distic.html?var_recherche=DISTIC).

les travaux du Loria liés aux besoins d'information dans un contexte de prise de décision : David-2005, Goria 2006, Kislin 2007).

Nous proposons, dans ce travail, de mettre en place un dispositif socio-technique d'infomédiation entre offre et demande d'information, qui se réfère à la théorie des facettes. Ce dispositif vise à proposer plusieurs grilles de lecture possibles d'un univers informationnel complexe (Morin et Lemoigne, 2007). Il est intéressant de re-situer dans un premier temps ce projet dans un contexte plus large. Nous proposons donc un détour par l'analyse de données, la philosophie, l'art, les sciences de gestion, l'informatique pour faire ressortir des concepts de multidimensionnalité, de « points de vue », d'infomédiation, de perspectivisme qui, tout en ayant été développés chacun dans des contextes différents n'en réfèrent pas moins à la même réalité. Ces concepts fournissent plusieurs vues permettant de resituer la théorie des facettes, qui nous intéressera ici, dans un contexte plus général.

L'esprit cartésien repose sur le principe de décomposition des phénomènes appelée disjonction par Morin (1982) et considère que la somme des parties constitue le tout.

La question de l'angle à choisir pour représenter un phénomène complexe est au cœur de l'analyse de données. Il s'agit alors de projeter une réalité multidimensionnelle sur un espace réduit qui puisse refléter au mieux la diversité du phénomène à décrire. La comparaison prise par Fénelon (1981) dans son ouvrage d'analyse de données est éclairante de ce point de vue : l'auteur prend l'exemple d'un chameau dont on souhaiterait rendre compte sur un plan. Plusieurs coupes du chameau peuvent être envisagées mais celle qui rend le mieux compte de la spécificité de l'animal est sans doute celle de profil. Cette recherche de la meilleure vue possible correspond à la solution qui restitue le maximum d'information contenue dans les données brutes.

Pour Leibniz, l'univers est une série infinie de courbes. Deleuze (1986) construit la notion de « point de vue » à partir des travaux de Leibniz. Le « point de vue » se loge dans chaque courbe. Cette vision renferme la notion de pluralité de « points de vue ». Le « point de vue » n'est donc pas une perspective frontale qui permettrait de rendre compte d'une forme de façon optimale mais une perspective locale. Pour Monnoyer Smith (2008), « Aucune

vision surplombante n'étant à portée humaine, la multiplication des points de vue constitue une richesse pour toute communauté d'individus ». Il n'y a pas un seul point de vue auquel il faudrait se hisser mais plusieurs points de vue qui permettent à chacun, dans une logique perspectiviste, d'organiser le chaos. Chacun n'éclaire qu'une portion et a des degrés de conscience plus ou moins claire, plus ou moins obscure, plus ou moins confus.

Des travaux en informatique s'intéressent à la combinaison de l'approche locale et globale. Les techniques de visualisation « focus + contexte » (Sarkar et Brown 1992, Furnas 1986) proposent à l'utilisateur de combiner une vue générale et un focus sur un sous ensemble. Le niveau de détail de l'élément à afficher dépend de sa distance au focus : suivant le focus choisi, certains éléments peuvent être détaillés, simplifiés ou simplement supprimés. Ainsi on obtient des représentations qui combinent une représentation macroscopique et une représentation autour d'un focus. La Figure 4 fournit un exemple de distorsion géométrique.

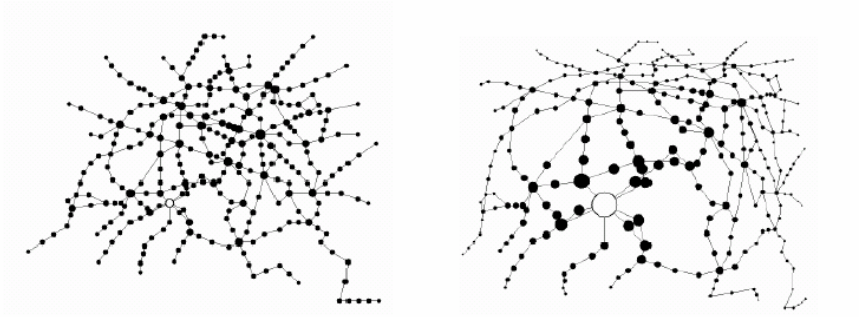


Figure 4 : ***application d'un zoom géométrique***

Dans l'art contemporain, les artistes utilisent plutôt des points de vue simultanés pour montrer la complexité des choses. L'œuvre de Joseph Kosuth, *One and three hammers* (1965) présentée Figure 5, en fournit un exemple en juxtaposant un marteau, une photo de ce marteau et la définition du dictionnaire d'un marteau.

Figure 5 : ***Joseph Kosuth, One and three hammers (1965)***

Dans un contexte d'intelligence économique, les travaux du Loria (David-2005, Goria–2006, Kislin- 2007) s'intéressent à la fonction d'infomédiaire dont le rôle est de traduire un besoin d'information exprimé par un décideur en un problème de recherche d'information.

En guise de synthèse, on peut donc voir que « pour étudier le monde réel et ses phénomènes, il est nécessaire d'en élaborer une représentation » (Rousset el al., 2005). Cette représentation est selon le cas une vue globale ou partielle, unique ou multiple. La théorie des facettes que nous allons développer consiste précisément à présenter plusieurs vues partielles sur un corpus documentaire et à proposer à l'internaute de combiner ces vues en fonction de ses préférences pour se construire sa vue à lui.

B. entrer dans la perception de l'usager

Manon est en classe de CM1. Sa maîtresse a demandé à chaque enfant de la classe de préparer un exposé sur un sujet du programme de l'année. Manon a choisi « l'énergie nucléaire » … et toute la famille est mobilisée. Une première collecte de données est réalisée sur *Google*. Il s'agit de trouver des pages web de vulgarisation sur le sujet compréhensibles pour cette enfant de 9 ans. Quelle requête[14] faut-il poser à *Google* pour obtenir le résultat recherché? On a tenté

[14] La requête désigne la question adressée par l'internaute au moteur de recherche.

« énergie nucléaire vulgarisation » mais les pages web de vulgarisation sur « énergie nucléaire » ne comportent pas nécessairement le terme « vulgarisation ». Manon et sa famille sont réduits, faute de mieux, à explorer une à une les réponses à une requête plus générale « énergie nucléaire ». La 125ème réponse de *Google* trouvée, au bout de 15 minutes de navigation, semble un point de départ intéressant. Pour effectuer cette recherche nous avons dû privilégier la découverte du corpus par la navigation (*browsing*) plutôt que par la recherche (*searching*).

Franck travaille pour la société Areva. Il doit réaliser pour le compte de sa société un dossier d'information stratégique pour identifier les sites web dont le contenu est susceptible de constituer une attaque informationnelle pour la société qui l'emploie. Dans le cadre de cette « veille d'image »[15], Franck cherche à identifier des pages web dont le contenu textuel ou figuratif est négatif et porte atteinte aux intérêts de sa société. Franck a beaucoup de mal à formaliser la bonne requête pour interroger *Google*. La requête « attaque informationnelle Areva » ne donne pas de résultats pertinents. Il s'essaie alors à des requêtes plus spécifiques comme « Areva scandale» qui donne des résultats plus intéressants mais toutes les pages critiques sur Areva ne comportent pas le terme « scandale ». Bref, cette requête présente peu de bruit[16] mais un grand silence[17]. Or Franck ne peut pas se permettre de laisser passer une page négative.

Christine est déficiente visuelle. Il existe une obligation légale faite aux sites web publics de produire des contenus accessibles mais combien d'entre eux la respectent ? Une vingtaine de sites web en France disposent du label Accessiweb. Christine possède un appareil auditif lui permettant d'entendre le

[15] Nous désignons par « veille d'image » une démarche de collecte et d'analyse humaine et/ou automatique de l'information dont l'objectif est d'identifier le plus en amont possible des signaux faibles de nature, s'ils ne sont pas perçus à temps, à dégrader l'image de l'organisation.

[16] Dans le contexte de la recherche d'information sur internet, le bruit correspond à des documents renvoyés par le moteur de recherche qui ne correspondent pas à la requête. Ces documents sont donc hors sujet.

[17] Dans le contexte de la recherche d'information sur internet, le silence correspond à des documents qui correspondraient à la requête mais qui ne sont pas renvoyés par le moteur de recherche

contenu textuel des pages web ainsi que le contenu des balises de texte lorsqu'il y a une image. Mais Christine perd beaucoup de temps quand elle est sur Internet : une mauvaise structuration du document rend sa recherche fastidieuse. Elle aimerait disposer d'un indicateur qui lui permette de mesurer, pour une page web donnée, le niveau d'accessibilité de la page web qu'elle s'apprête à parcourir.

Pedro travaille dans une société qui développe des outils de coupe utilisés dans le forage pétrolier off shore. Il est chargé d'estimer le marché des diamants poly-cristallins au Brésil. Avant d'aller consulter les experts du domaine, il aimerait disposer d'une première estimation à partir de sources provenant du web. Pedro se souvient que la valeur d'une chaîne est celle de son maillon le plus faible. La valeur des résultats qu'il obtiendra dépendra de la validité des données qu'il aura collectées en amont mais aussi de leur pouvoir couvrant par rapport à l'ensemble des informations sur le sujet. A défaut de solution systématique satisfaisante, Pédro adresse à *Google* la requête « poly-cristallins diamond filetype :gov » : il sait que les informations présentes sur les sites gouvernementaux sont validées en amont et qu'il pourra les exploiter. Il y a sans doute d'autres informations sur le web qui pourraient être utilisées mais Pedro préfère ne pas prendre de risque en introduisant, dans son étude, des données qui pourraient provenir d'acteurs non compétents ou mal intentionnés.

C. Problématique

Manon, Christine, Franck et Pedro ont un point commun : leur recherche d'information ne peut pas être correctement traduite sous forme de mots clés. La requête qu'ils adressent au moteur de recherche ne permet de définir que le sujet (*topic*) de leur recherche. Une recherche d'information pourrait être définie par des dimensions complémentaires comme par exemple son niveau de subjectivité[18] (utile pour Pédro), son niveau de lisibilité (utile pour Manon), son niveau d'accessibilité (utile pour Christine), son orientation négative (utile

[18] les choses sont présentées pour l'instant de façon intuitive. Les notions sous jacentes (subjectivité, lisibilité…) feront l'objet d'un cadrage rigoureux dans les pages qui suivent.

pour Franck). Ces dimensions ne peuvent pas être exprimées dans la requête adressée à un moteur de recherche. Tous les internautes n'auront pas besoin de mobiliser chacune de ces dimensions complémentaires, chacune correspondant à un besoin très précis. Cette intuition d'un nécessaire enrichissement de la requête par des dimensions complémentaires doit être maintenant précisée pour en comprendre les raisons et les modalités d'implémentation.

Pour quelles raisons nos 4 internautes sont-ils amenés à suggérer ces dimensions complémentaires dans leur recherche d'information ? La réponse est à rechercher dans la spécificité du marché de l'information sur internet. Le monde du web se caractérise en effet par une grande diversité de l'information offerte et par la diversité des demandes exprimées par les internautes. Illustrons chacun de ces éléments par un exemple.

Lorsqu'on effectue une recherche d'information dans une base de brevets, on sait implicitement qu'on dispose d'une source d'information fiable, validée en amont. L'information généraliste sur le web n'est validée par aucune autorité et par aucune chaîne éditoriale. Il faut donc, pour pouvoir utiliser l'information disponible sur internet, dans un cadre professionnel, disposer d'heuristiques permettant de distinguer l'information valide de celle qui ne l'est pas.

Lorsqu'un enfant recherche un ouvrage dans le rayonnage pour enfant d'une bibliothèque, il sait qu'il trouvera un livre de son âge. Lorsque le web est sollicité par des internautes ayant des âges et des besoins différents, il est naturel de chercher à spécifier plus avant une information correspondant à ses préférences.

Si on veut pouvoir répondre à Manon, Christine, Franck et Pedro, il faut enrichir les données d'une page web par des métadonnées. Plusieurs solutions ont été envisagées dans les travaux antérieurs pour y parvenir : recourir à des experts du domaine, au concepteur de la page web, aux internautes ou mobiliser des routines automatiques.

Dans la logique de la « recherche d'information bibliographique », un certain nombre de tâches de classification et de validation de l'information sont réalisées en amont par des professionnels de la documentation au bénéfice de la communauté des utilisateurs de ces systèmes[19]. Dans la logique des moteurs de recherche, cette tâche de validation et de classification, au lieu d'être assurée par une autorité en amont, est déportée sur l'utilisateur final et a lieu à chaque fois que la ressource d'information est utilisée. Pour Davis (2005), le passage d'une classification contrôlée à une classification non contrôlée (de type web) conduit à un transfert de temps de l'indexeur à l'utilisateur. En terme macroéconomique, le traitement en amont est peut-être plus efficace car il n'a lieu qu'une seule fois. Toutefois, l'organisation décentralisée du web et sa croissance exponentielle sont-elles encore compatibles avec le modèle de la validation amont ?

La deuxième façon d'enrichir le contenu d'une page web par des métadonnées consiste à demander au concepteur de la page web de renseigner différentes rubriques relatives à cette page. Les travaux du World Wide Web Consortium (W3C) ont débouché sur l'établissement d'un standard (*Dublin Core*[20]) qui définit 13 métadonnées[21] caractérisant les pages web. Ces métadonnées doivent être saisies par les concepteurs de la page web. Dans la pratique, ce travail est rarement réalisé ce qui rend non significatifs les résultats issus de l'exploitation de ces données.

D. Notre approche

Notre recherche s'inscrit dans une logique de travaux qui visent à développer des indicateurs automatiques qui vont permettre de qualifier en amont une page web. Ces éléments sont alors proposés à l'internaute dans une interface *ad hoc* afin de lui permettre d'orienter sa recherche.

[19] C'est le cas des annuaires de recherche qui emploient des professionnels chargés de sélectionner l'information pertinente sur un sujet
[20] http://dublincore.org/
[21] http://www.dublincore.org/documents/dces/

Pour bien montrer au lecteur la finalité de ce travail, nous avons conçu un démonstrateur[22]. Ce démonstarteur a été construit à partir de l'outil Facetmap [23] dont une version gratuite est disponible sur internet. Ce genre d'outil est utilisé par de nombreux sites de e-commerce qui sont de grands utilisateurs de théorie des facettes. La *Figure 6* présente une capture d'écran de ce démonstrateur. Lorsqu'il effectue une recherche d'information sur un sujet, l'internaute adresse une requête au moteur de recherche (dans le cas de cet exemple « énergie nucléaire » comme aurait pu l'adresser Manon ou Franck).

Figure 6 : exemple de résultat du démonstrateur

Les réponses à la requête sont alors catégorisées automatiquement autour de plusieurs dimensions. Dans le cas de la *Figure 6*, nous avons retenu 5 dimensions (*valence*, *web page objectivity level*, *web page readibility*, *web page accessibility*, *web site trafic*). En fonction de son besoin d'information (l'internaute a fait ici le choix de privilégier les pages à valence[24] négative),

[22] Le démonstrateur est un site web qui permet de représenter l'interface du moteur de recherche à facettes que nous proposons. Cette interface permet une navigation au sein d'un corpus prétraité de données. Les indicateurs caractérisant chaque facette ne sont donc pas définis en temps réel.

[23] http://facetmap.com/demosetup/

[24] Typiquement « bon » a une valence positive et « mal » une valence négative

l'internaute affine sa recherche et les résultats à l'affichage se mettent à jour automatiquement (39 résultats satisfont à sa recherche). La recherche d'information s'effectue par navigation non plus seulement dans l'univers des pages web mais dans des dimensions qui correspondent chacune à un éclairage apporté sur le corpus à présenter.

Ce démonstrateur malgré son caractère intelligible et séduisant masque différentes questions auxquelles il nous faudra répondre.

- A quel courant théorique peut-on rattacher ce travail de terrain ? Nous serons conduits à préciser que ce travail s'inscrit dans le courant théorique de la « théorie des facettes » développé par Ranganthan (1933).
- Pour ne prendre que les exemples qui sont donnés, les dimensions de « valence » ou de « niveau de subjectivité » d'une page web sont-elles réductibles à un indicateur automatique ? Pour répondre à cette question, nous serons amenés, pour chaque dimension, à réaliser un état de l'art qui nous permettra d'explorer des travaux antérieurs dans le domaine de la linguistique computationnelle et de la psychologie appliquée. A partir de ces travaux, nous établirons pour chaque dimension un indicateur automatique. Pour tester la pertinence de chaque indicateur, nous considérerons un corpus de pages web. L'indicateur automatique sera calculé sur chaque page web de ce corpus. Chaque page du corpus sera par ailleurs soumise à une qualification humaine en aveugle. L'indicateur automatique sera alors comparé à l'évaluation faite par des juges humains. La pertinence de nos indicateurs sera appréciée par la concordance mathématique entre jugement humain et jugement automatique.
- Quelles sont les différentes dimensions que nous pourrions concevoir dans une telle interface ? La théorie des facettes, que nous utilisons, est ouverte en ce sens qu'elle permet d'intégrer de nouvelles facettes.
- L'utilisateur de l'interface à facettes a-t-il la possibilité de sélectionner et de combiner entre elles les facettes de son choix pour se constituer son propre indicateur de pertinence ?

- Cette interface de moteur de recherche est-elle innovante ? Un détour par la pratique nous permettra de constater qu'on retrouve ce genre d'interface dans des sites web marchands. De ce point de vue, notre travail s'apparente à une transposition de ces initiatives au monde la recherche d'information. Ce travail est plutôt une systématisation d'une démarche qu'on observe dans le monde des moteurs de recherche sous la forme de moteurs à curseurs.
- Peut-on juger de l'efficacité de ce genre d'interface ? S'avère-t-elle plus efficace, plus rapide qu'une interface de recherche traditionnelle pour rechercher de l'information ? La réponse à ce genre de question suppose que l'on dispose d'un prototype interrogeable par l'internaute en temps réel. Nous n'en sommes pas arrivés jusque là. Nous disposons d'un démonstrateur permettant pour un petit nombre de requêtes de classer les résultats autour de certaines dimensions.

Pour répondre à ces différentes questions, nous allons organiser la réflexion autour d'un plan en plusieurs chapitres :

Le premier chapitre a pour objectif de présenter notre démarche comme évolution naturelle de la dynamique qui s'opère depuis 10 ans dans le monde de la recherche d'information sur Internet. Nous identifierons les défis que doivent relever les moteurs de recherche en matière de personnalisation des résultats. Nous serons alors à même de montrer que notre approche est une solution possible à ces défis.

Le second chapitre présentera la méthodologie que nous avons retenue dans ce travail de recherche.

Au chapitre 3, nous envisagerons les fondements théoriques de cette recherche à travers des développements relatifs à la théorie des facettes développée par Ranganathan. Il s'agira tout d'abord de préciser le terme de facette que l'usage courant et la complexité des travaux de Ranganathan n'ont pas toujours permis de clarifier. Cette théorie a été développée dans les années 1930 à l'heure où le web n'existait pas. Il sera donc nécessaire de re-situer les travaux des pères

fondateurs et les prolongements qu'ils ont pu connaître dans l'ère numérique. Cette partie permettra de voir que notre démarche s'inscrit dans une définition extensive de la théorie des facettes et en constitue un prolongement original que nous n'avons pas retrouvé par ailleurs.

Le chapitre 4 consistera à décrire les facettes que nous avons retenues. Pour chaque dimension, une démarche expérimentale sera conduite pour comparer, sur un jeu de données, les résultats de l'indicateur automatique et les résultats donnés par des juges.

Le dernier chapitre présentera quelques réflexions sur l'implémentation possible de ces facettes dans une nouvelle interface de moteur de recherche.

CHAP1 : UN CONTEXTE DE TERRAIN FAVORABLE

L'objet de ce chapitre est de montrer que ce sujet est en phase avec les évolutions récentes du monde des moteurs de recherche. Nous allons commencer par présenter une analyse critique du secteur de la recherche d'information sur Internet. Nous montrerons que les moteurs de recherche majeurs sont réducteurs par la linéarité des résultats qu'ils proposent, par leur faible nombre de résultats exploitables et par leur instabilité. Leurs indicateurs de pertinence reposent sur une certaine vision du monde et sur des algorithmes monolithiques et opaques pour l'usager. La concentration des acteurs du monde des outils de recherche conduit à des barrières à l'entrée qui rendent l'arrivée de nouveaux entrants sur le marché difficile : la structure oligopolistique du marché est un facteur de forte réduction de la diversité et de l'innovation. Tous ces facteurs nous conduisent à remettre en cause la confiance de principe qu'a un usager dans le résultat des moteurs de recherche et l'exploitation sans recul des résultats des moteurs de recherche dans un contexte professionnel. En partant de cette vision critique, nous pourrons mieux comprendre les perfectionnements des moteurs de recherche dans le sens d'un indicateur de pertinence orienté usager et mettre en lumière l'avantage de l'approche proposée par la théorie des facettes.

A. Analyse critique des moteurs de recherche

Sans nier l'apport fantastique des moteurs de recherche, nous proposons dans cette partie une analyse critique de leurs méthodes et résultats en nous concentrant plus particulièrement sur Google. Ce choix s'impose par la dominance de Google sur le marché des moteurs de recherche et se justifie

aussi par le fait que son indicateur de pertinence est très proche de celui des moteurs concurrents. Les autres moteurs ne seront cités que dans la mesure où ils se distinguent de ce modèle général.

1. Vision linéaire d'un corpus web en interaction

Les réponses à un moteur de recherche sont souvent présentées sous forme de listes de pages web offrant ainsi un ensemble de réponses disjointes. La logique relationnelle, si elle est introduite dans les algorithmes de pertinence des moteurs de recherche, n'est pas utilisée par les moteurs majeurs dans la restitution de cette information. Il en découle qu'il est difficile d'avoir une vision d'ensemble d'un corpus web alors même que le lien hypertexte est constitutif du web au même titre que la page web elle même.

2. Vision fragmentaire de l'abondance

Les moteurs de recherche actuels exploitent un nombre énorme de données. Dans tous les cas, seules les 1000 premières pages web seront consultables par l'internaute. Cet échantillon est-il représentatif de l'ensemble des réponses pouvant être proposées par le moteur de recherche ? L'étude des usages montre aussi que l'internaute se satisfait d'un très petit nombre de réponses (il analyse les 20 premiers liens maximum de la liste qui lui est fournie dans 80% des cas). Dans une démarche de recherche d'information non professionnelle, on peut se satisfaire d'un tel ensemble de données. Cela est plus difficilement acceptable dans le cadre d'une recherche d'information professionnelle. L'internaute professionnel accorde parfois une confiance excessive à des conclusions issues de l'exploitation d'informations obtenues à partir d'échantillons non représentatifs. C'est ce que Tversky et Kahneman (1974) appellent « la loi des petits nombres », qui est une forme de biais cognitif observé dans le processus d'appréhension d'une information massive.

3. Des algorithmes de moins en moins lisibles

Plusieurs familles d'indicateurs de pertinence sont mobilisées par les moteurs de recherche. On distingue les indicateurs reposant sur l'analyse du contenu textuel de la page web (*content oriented),* ceux qui utilisent le lien hypertexte comme source de légitimité par la page qui le reçoit *(link oriented),* ceux qui

sont centrés sur l'usager *(user oriented)* et ceux qui privilégient une logique marchande *(business oriented*). Bien souvent les indicateurs utilisés par les moteurs de recherche combinent ces différentes approches de façon subtile.

Le critère permettant de juger de la qualité d'un indicateur est son caractère exogène aux objets mesurés. Si la pertinence d'une page web pour un mot clé donné est mesurée uniquement par le nombre de fois où ce mot clé est présent dans la page web, on est en présence d'un indicateur dans lequel la page web elle même peut être actrice de sa pertinence. On parle dans ce cas d'indicateur de pertinence endogène. L'évolution de la recherche dans le domaine des indicateurs de pertinence a donc consisté à trouver des indicateurs de pertinence d'une page web qui ne dépendent pas de la page elle-même mais qui lui soient le plus possible extérieurs. *Google* a fondé sa puissance sur son algorithme de PageRank (Brin S, Page L., 1998) qui considère qu'une page web tire sa pertinence de la légitimité qui lui est conférée par les pages web qui la citent. Le PageRank de *Google* repose donc, sur le papier, sur un indicateur de pertinence exogène. Toutefois, l'analyse relationnelle du PageRank sous jacente au modèle de *Google* (et dont le principe est généralisé depuis aux autres moteurs de recherche) est de moins en moins exogène comme l'on montré les références au cyclotron[25] dans la littérature spécialisée.

Aujourd'hui le référencement est devenu un métier car les enjeux commerciaux d'un positionnement dans les premières réponses d'un moteur de recherche sont importants. Il y a une course entre moteurs et référenceurs. Les premiers proposent un indicateur censé présenter les pages web selon leur pertinence « objective » par rapport à la question posée. Les seconds cherchent le moyen de contourner cet indicateur pour être bien positionnés dans les résultats du

[25] Au sens de *Google*, la pertinence d'une page web est définie par la pertinence des pages web qui pointent sur elle. La pertinence de ces pages web pointantes est définie de la même manière. Il est possible pour une page web A de créer un cycle vertueux qui consiste pour la page A à créer un lien hypertexte vers une page B qui elle-même possède un lien hypertexte vers une page C qui possède un lien hypertexte vers la page A. Grâce à l'existence de ce cycle, la page devient indirectement actrice de sa propre pertinence. Ce problème appelé « cyclotron problem » a conduit à des pratiques de contournement pour améliorer le référencement d'un site sur le web.

moteur de recherche. Aujourd'hui les moteurs de recherche répondent à cette compétition de différentes manières :

- Par la sanction : les principaux moteurs de recherche créent des listes noires comportant les sites accusés de détourner à leur profit l'algorithme de pertinence des moteurs de recherche. Il en va ainsi des sites qui surchargent leurs balises méta de mots clé redondants, des pages qui affichent un texte de la même couleur que le fond, de pages web qui pratiquent le *cloaking*[26]. Au fur et à mesure que le temps passe, on observe une surenchère et des techniques de contournement de plus en plus sophistiquées.
- Par la discrétion. *Google* s'est rendu compte en publiant son brevet PageRank que la protection industrielle que lui conférait ce brevet était une arme à double tranchant puisqu'elle donnait à l'internaute et au référenceur les moyens de connaître l'indicateur de pertinence. Les contournements n'ont d'ailleurs pas tardé. Les moteurs ont aujourd'hui changé de stratégie et communiquent très peu sur leur indicateur de pertinence. Actuellement *Google* utilise des algorithmes complémentaires dans son indicateur de pertinence qui reposent sur la localisation géographique[27], le contenu du texte figurant sur les liens entrants[28] sans que le poids de chacun de ces éléments soit clairement explicité. Cette discrétion soulignée par Diaz (2005) est extrêmement gênante pour l'internaute qui se voit proposer une vision du monde suite à une requête qui dépend de la vision du moteur sans que celle-ci soit clairement explicitée.

[26] une page web "cloakée" existe en deux versions : une destinée aux robots des moteurs de recherche et une destinée à l'utilisateur final. L'intérêt est d'optimiser une page web pour satisfaire au mieux aux critères pris en compte par le moteur de recherche pour définir son indicateur de pertinence. Dès que le robot du moteur de recherche est passé pour mettre à jour son index, la page est remplacée par la page qui sera visualisée par l'internaute.

[27] La localisation géographique permet notamment au moteur d'exclure certaines pages qui ne respectent pas la réglementation de certains pays. Elle permet également l'affichage de publicités ciblées sur la localisation de l'internaute.

[28] La pratique du *Google bombing* a rendu cet algorithme populaire. Il consiste à améliorer le classement d'une page web (notée A) pour une requête donnée en indiquant le texte de la requête sur plusieurs liens pointant sur la page web A. C'est ainsi que la requête « arme de destruction massive » sous *Google* a pu renvoyer à une page « page not found ». Le texte « arme de destruction massive » constituait le texte de plusieurs liens hypertextes qui pointaient tous vers une page qui n'était volontairement plus accessible.

- Par le changement : les moteurs modifient régulièrement leurs indicateurs de pertinence ou du moins le poids accordé, au sein de l'indicateur de pertinence, à telle ou telle dimension, rendant plus difficile encore le positionnement artificiel d'une page web.

4. Vision subjective du monde

Le modèle *Google* qui est actuellement le modèle de référence et que suivent les autres outils de recherche majeurs (MSN, Yahoo !) repose sur des hypothèses sous jacentes rarement formulées qui mériteraient d'être explicitées. Ce modèle repose sur un algorithme mathématique qui définit la pertinence sur la base des liens hypertextes pointant sur une page web. Cette vision des choses fonctionne bien dans un système où le lien hypertexte signifie la reconnaissance par le citant de la page citée. Avec le développement du web marchand, le lien hypertexte s'éloigne de cette signification originelle. Il découle aussi de cette vision relationnelle que les moteurs de recherche favorisent dans leur classement les pages web anciennes au détriment des pages émergentes qui ont une probabilité plus faible, ceteris paribus, d'avoir un taux de citation égal à celui de pages plus anciennes. Othon et al. (2004) parlent de biais temporel pour désigner ce phénomène. Un biais est une tendance du moteur de recherche à valoriser certaines pages web pour des raisons non intentionnelles (le cas ici car l'algorithme de pertinence, reposant sur un critère relationnel, va valoriser les pages anciennes) ou intentionnelles (logique marchande où il faut payer pour apparaître dans les premières réponses du moteur de recherche). Baeza-Yates et al (2002), ont également étudié le biais du PageRank vis-à-vis des nouvelles pages.

5. *Google*archie ou loi du plus fortement relié

Ce terme développé par Hindman et al. (2003) part du constat établi par Barabási et Albert (1999) et Broder et al (2000) selon lequel le nombre de liens entrants sur une page web suit une loi de distribution de type puissance. Cela signifie qu'un petit nombre de pages web reçoit le plus grand nombre de liens entrants. De par la nature relationnelle de l'indicateur de pertinence des moteurs de recherche, ces pages sont les mieux classées dans les résultats du moteur, donc les plus visitées par les internautes, donc les plus susceptibles

d'être pointées comme page de référence. Cela a tendance à renforcer leur classement dans le moteur de recherche. Cho et Roy (2004) ont montré que les pages les plus populaires devenaient de plus en plus populaires et celles qui ne le sont pas le sont de moins en moins. Diaz (2005) ne dit pas autre chose lorsqu'il précise que des algorithmes de type PageRank donnent un porte-voix aux pages les plus puissantes et une muselière aux pages sous représentées. Toujours selon Diaz, PageRank favorise le statu quo. Or le progrès passe par la subversion du statu quo.

Il serait intéressant de proposer un outil de recherche qui puisse opérer un brassage des réponses du moteur de recherche et non une concentration des mêmes pages dans le haut du classement.

6. *Google*ocratie ou la démocratie en question.

Un des principes d'une démocratie idéale est « une personne une voix ». On peut présenter le classement des pages de *Google* comme un système de vote dans lequel les préférences exprimées par les auteurs de pages web sont agrégées. Les auteurs de pages web votent en créant des hyperliens et *Google* compte les votes. D'un point de vue macroscopique il y a démocratie puisque le PageRank ne fait qu'agréger ce que des millions d'auteurs de pages web ont jugé important. On peut observer par ailleurs une corrélation forte entre le nombre de liens entrants sur une page web et le nombre de visiteurs qu'elle recevra. Cette corrélation renforce l'idée selon laquelle *Google* agrège finalement des données permettant de donner du poids à l'expression des concepteurs et des utilisateurs du web. Toutefois, cette idée de démocratie ne résiste pas à une analyse plus fine comme l'a montré Diaz (2005). En effet, *Google* enfreint la règle « une personne une voix » puisque d'après le PageRank, une page sera d'autant plus pertinente qu'elle sera citée par des pages pertinentes. Les pages pertinentes ont donc un poids plus fort que les autres.

7. Googleopole ou absence de diversité

Les outils de recherche laissent de moins en moins la diversité s'épanouir et connaissent une homogénéité. Le secteur des outils de recherche est un secteur qui devient de plus en plus concentré entre les mains de quelques acteurs. Diaz

(2005) parle de « *Google*opole ». Les évolutions des dernières années voient apparaître trois acteurs majeurs qui se sont constitués pour certains par croissance externe (Yahoo ![29]) et d'autres principalement par croissance interne (MSN[30], *Google*[31]). Ces acteurs ont un indicateur de pertinence constitué à partir d'une base technologique très voisine pour autant qu'on puisse avoir des informations suffisantes pour pouvoir en juger. En 2008, ces trois acteurs représentent environ 80 à 90% de la recherche d'information tant en France (*Tableau 3*) qu'au niveau mondial (*Tableau 4*) [32]

TOP 15 des Moteurs de recherche		
Part de visites des familles de moteurs	*Janvier 2008*	*Février 2008*
1. Google	90,03%	89,97%
2. Yahoo!	2,88%	3,03%
3. Live Search	2,49%	2,54%
4. AOL	1,54%	1,55%
5. Orange	1,56%	1,52%
6. Free	0,48%	0,46%
7. Alice	0,36%	0,35%
8. Altavista	0,08%	0,07%
9. Virgilio	0,08%	0,07%
10. Ask	0,08%	0,06%
11. Voilà	0,06%	0,06%
12. Exalead	0,06%	0,06%
13. Seznam	0,05%	0,04%
14. Lycos	0,04%	0,03%
15. MySearch	0,03%	0,03%

Tableau 3 : Statistique établie par XiTi du 1er au 29 février 2008, à partir des statistiques de trafic de 89621 sites web francophones[33]

[29] http://www.yahoo.com
[30] http://www.msn.com
[31] http://www.*Google*.com
[32] statistiques fournies par Xiti
[33] *http://www.webrankinfo.com/actualites/200803-parts-marche-moteurs-france-fevrier-2008.htm*

Core Search Entity	November 2007 (%)	December 2007 (%)
Total core search	100.0	100.0
Google sites	58.6	58.4
Yahoo sites	22.4	22.9
Microsoft sites	9.8	9.8
Time Warner network	4.5	4.6
Ask network	4.6	4.3

Tableau 4: Statistique établie par comScore Décembre 2007[34],

Il existe sur le marché de la recherche d'information de nombreux petits outils de recherche mais ils peuvent difficilement se présenter comme des alternatives sérieuses aux moteurs majeurs de par leur faible couverture du web. Ces outils n'ont souvent pas de base de données propre et lorsqu'ils en ont une, elle n'est pas de taille suffisante. Ils offrent souvent une vitrine de leur technologie à partir de corpus de données récupérés d'annuaires gratuits (type Open Directory Project[35]) mais qui comportent quelques millions de pages seulement. Ils interrogent parfois les moteurs de recherche majeurs en tant que méta moteurs dépendant alors de la bonne volonté du moteur de recherche sollicité qui peut très bien du jour au lendemain bloquer les vannes ou renvoyer des informations mal classées. Ces outils pourraient servir de complément aux moteurs majeurs du fait de l'originalité technologique qui les fondent. Dans la réalité, ils peuvent difficilement apparaître comme des sources d'information sérieuses pour le professionnel. On a aujourd'hui une asymétrie du contexte et des barrières à l'entrée telles qu'il est difficile d'envisager à court terme l'arrivée d'un nouvel entrant sur ce marché. En effet, être acteur dans le monde des outils de recherche représente un investissement qui se compte en centaines de millions de dollars. Les acteurs en présence ont réussi par la protection industrielle que leur confère leur brevet à fermer le secteur aux nouveaux entrants.

[34] *http://searchenginewatch.com/showPage.html?page=3628341*
[35] http://www.odp.com

8. Manque de stabilité des résultats

Le contenu du web change et il est normal que l'index des moteurs de recherche suive cette évolution. Des statistiques macroscopiques réalisées en 2004 par Ntoulas et al (2004) montrent que 8% de nouvelles pages web sont créées chaque semaine. Il est important que les index des outils de recherche se rafraîchissent pour renvoyer une information la plus en phase avec la réalité du moment. La question de la stabilité des réponses des outils de recherche ne se situe pas à ce niveau.

Ce qui préoccupe, c'est le fait qu'un moteur de recherche interrogé au même moment peut renvoyer, en réponse à la même requête, un nombre de réponses sensiblement différent. Pinczon du Sel (2006), s'est intéressé à cette question et a testé la stabilité du moteur de recherche Google. L'objectif est de mesurer les écarts dans les résultats de ce moteur en se soumettant à un protocole strict d'interrogation du moteur Google à une période qui ne se caractérise pas par le phénomène de Google Dance[36].

Google dispose de différentes bases d'index (data centers) dispersées dans le monde[37] et interrogées de façon transparente par l'utilisateur suivant sa localisation.

Lors d'une expérimentation, nous avons comparé les résultats donnés par trois de ces data centers ayant pour adresse IP : 64.233.183.104, 66.102.9.104 et 216.239.59.104. Un jeu de requêtes a été soumis à ces serveurs afin de voir si leurs bases de données étaient synchronisées. Nous avons observé des résultats souvent différents. Cette différence se situe à un double niveau :

- Le nombre total de réponses renvoyées par le moteur *Google* est très différent selon le data center sollicité. L'écart maximal observé est noté en pourcentage dans le *Tableau 5* pour diverses requêtes. Par exemple pour la requête « the bachelor », on note un écart de 30,88 % entre le nombre de résultats retournés par les data centers 66.102.9.104 (ayant

[36] Période de recalcul du PageRank qui se traduit par une absence ponctuelle de synchronisation des index du moteur de recherche *Google*

[37] Il est possible d'interroger Google sur plusieurs de ces bases distinctes : http://www.webrankinfo.com/outils/google-dance/google-dance3.php. Le 20 Mai 2008, nous avons obtenu 10 data centers actifs

48 900 000 réponses) et le datacenter 64.233.183.104 (33 800 000 réponses).

Mots-clés	Ecart maximal en % entre data centers
the bachelor	30,88
e3	56,26
kasey kahne	54,85
kylie minogue	63,43
natalie portman	36,44
nintendo revolution	99,79
preakness	16,83
ps3	20,45
star wars episode 3	37,05
xbox 360	48,05
brice de nice	34,66
caf	50,24
impôts	49,14
jeux	80,37
manga	73,51
meetic	0,89
pape	52,60
pmu	29,86
sfr	63,48
tiscali	8,99

Tableau 5 : Ecart maximal en pourcentage entre 3 data centers de Google

- Le classement des pages web renvoyées diffère entre les trois data centers. Pour étudier ce point, nous avons noté la position du premier changement de rang entre les classements proposés par chaque index. Nous avons limité l'analyse aux 100 premières réponses du moteur. Disposant de trois data centers pour google.com, le nombre de réponse de chaque data center a été comparé à l'index de référence (66.102.9.104). Les résultats sont présentés *Tableau* 6; La valeur de 28 dans la première ligne signifie que pour la requête « the bachelor », les 27 premières réponses du data center 216.239.59.104 et 66.102.9.104 sont identiques : le classement de ces deux data center est modifié à partir de la réponse 28. Quelques enseignements peuvent être tirés de ce tableau. Le classement des 20 premières réponses du moteur de recherche est modifié dans 40% des cas étudiés. Quand on sait que 80% des internautes se satisfont des 20 premières url de résultats, on

comprend que ces résultats ont un impact potentiel sur l'interprétation des résultats qui sera faite par l'internaute.

Mots-clés	216.239.59.104	64.233.183.104
the bachelor	28	25
e3	10	13
kasey kahne	22	19
kylie minogue	66	66
natalie portman	Idem	26
nintendo revolution	77	1
preakness	31	(N.C.)
ps3	22	2
star wars episode 3	9	10
xbox 360	Idem	13
brice de nice	8	6
caf	29	29
impôts	Idem	29
jeux	65	63
manga	Idem	11
meetic	58	9
pape	Idem	Idem
pmu	Idem	20
sfr	5	5
tiscali	38	6

Tableau 6: Position du premier changement de réponse par rapport au Google Data Center 66.102.9.104 (100 premières réponses)

Il est difficile de connaître les raisons de ces différences d'index : problème de synchronisation de bases de données géantes ou différences juridiques nationales impactant les index dans leur contenu ? Ce problème de stabilité peut avoir des conséquences préoccupantes surtout lorsque les résultats du moteur de recherche sont intégrés dans une chaîne de traitement de l'information qui nourrit un processus de veille. En effet dans le cas d'un usage professionnel de l'information provenant d'un outil de recherche, le problème de stabilité pose la question de la reproductibilité d'une recherche d'information et par là même de la scientificité d'une méthode qui reposerait sur une source d'information instable.

9. Vision imparfaite de l'interaction hypertextuelle

La clé de l'indicateur de pertinence des principaux outils de recherche est leur algorithme relationnel. Pour calculer leur indicateur de pertinence, les moteurs de recherche doivent disposer d'un index de pages web mais aussi du réseau des interactions qui existent entre ces pages web. La pertinence des résultats du

moteur de recherche dépend donc de la qualité du réseau des interactions hypertextuelles conservées par le moteur de recherche.

Des observations effectuées par Boutin et Perrin (2005-a) conduisent à une certaine remise en cause de la capacité du moteur *Google* à rendre compte fidèlement de l'interaction véritable entre des sites web donnés. Le travail décrit a été réalisé dans le cadre d'une étude expérimentale commandée par la Préfecture de région PACA et le Conseil Régional PACA où nous cherchions à mieux comprendre les interactions hypertextuelles entre les sites publics présents en région Provence Alpes Cote d'Azur. Nous avons donc constitué un corpus de sites web et nous nous sommes intéressés à la mesure de l'interaction entre les sites de ce corpus. Il s'agissait donc, sur un espace géographique cerné, d'identifier les interactions entre acteurs dont nous voulions rendre compte ensuite par des cartographies et des indicateurs. Pour collecter les liens hypertextes entre ces sites, nous avons utilisé deux techniques :

- La commande link disponible sur le moteur de recherche *Google identifie* pour une page ou un site web donné, toutes les autres pages qui pointent vers cette page ou ce site. Elle privilégie l'identification des liens entrants.
- L'outil Xenu's Link Sleuth (TM)[38] permet d'identifier les liens sortants pour une profondeur donnée d'un ensemble de pages ou de site web.

Les approches par liens entrants ou sortants sont différentes mais lorsqu'on s'intéresse aux liens existants au sein d'un ensemble circonscrit de pages web, ces deux méthodes doivent conduire logiquement aux mêmes résultats. Cependant, il semble, étrangement, que certaines pages web précédemment référencées par Google ne ressortent pas avec l'opérateur link alors qu'elles sont à l'origine d'un lien hypertexte entrant[39].

Ce type d'observation ne semble pas isolé puisque la méthode de recherche des liens sortants par Xenu nous a permis d'identifier environ deux fois plus de liens hypertextes que la méthode des liens entrants de *Google*. Il y a sans doute de bonnes raisons à cette discrimination des liens hypertextuels mais par simple observation, nous n'en avons pas trouvé. En tout état de cause, cela

[38] http://home.snafu.de/tilman/xenulink.html

[39] il suffit pour ça d'étudier les liens identifiés lors de la précédente Google Dance et présents dans le cache

pose un problème car *Google* présente des résultats en se basant sur un réseau d'interaction web qui n'est pas une représentation exhaustive et fidèle de la réalité des interactions hypertextuelles. Nous avons déduit de cette observation qu'il fallait être très prudent sur l'interprétation de la commande link de *Google* et qu'on pouvait avoir des réserves sur un algorithme relationnel qui opère sur un réseau d'interactions qui ne correspond pas aux interactions observables sur le web[40]. Des observations antérieures conduites par Bar-Ilan J (2002) ont conduit au même résultat.

B. Les défis actuels des moteurs de recherche

Ce travail de recherche a pour objectif de retracer et d'apporter une contribution à la double révolution que traversent actuellement les moteurs de recherche.

La première concerne le passage d'un indicateur de pertinence absolu à un indicateur de pertinence centré sur l'utilisateur. L'indicateur de pertinence est l'algorithme qui permet de hiérarchiser les pages renvoyées par le moteur de recherche. Classiquement les pages sont hiérarchisées les unes par rapport aux autres en fonction de critères définis par le moteur de recherche. Il est possible de penser des systèmes dans lesquels l'internaute attribuerait à certains critères des préférences particulières.

La seconde révolution traduit le passage d'une formulation de la requête intégrant exclusivement le sujet de la recherche (requête de type *topic related*) à une requête multicritères introduisant d'autres dimensions. Il ne s'agit alors plus de hiérarchiser les résultats d'une recherche les uns par rapport aux autres mais de filtrer les résultats obtenus pour n'afficher que ceux qui correspondent au souhait de l'utilisateur. Le *Tableau 7* retrace cette dynamique qui fait apparaître trois cases que nous allons maintenant détailler.

[40] L'étude a été réalisée sur Google. Un travail analogue mériterait d'être entrepris sur le moteur de recherche Yahoo !

		Type d'indicateur de pertinence	
		absolu	centré sur l'utilisateur
Type de requête	exprimant le sujet de recherche	*Moteur de recherche grand public (Google, Yahoo !)!*	I
	multicritères	II	III

Tableau 7: la double révolution des moteurs de recherche

Nous allons reprendre successivement ces deux dimensions.
Abordons dans un premier temps le glissement d'une pertinence absolue vers une pertinence relative.

1. Vers une pertinence relative

Si on a un regard diachronique sur la notion de pertinence au sens où l'entendent les moteurs de recherche, on s'aperçoit qu'on a connu depuis 10 ans plusieurs familles d'indicateurs de pertinence se combinant aujourd'hui:

- La première famille correspond aux indicateurs de pertinence basés sur l'analyse du contenu. Largement intrinsèque, elle définit la pertinence à partir de la présence du texte de la requête dans la page web à qualifier.
- La seconde famille, largement mise en œuvre et inspirée de *Google*, s'appuie sur l'analyse relationnelle. C'est alors l'étude des liens hypertextes qui permet de définir, de façon exogène, la pertinence d'une page web.
- La troisième famille est basée sur l'approche business retenue par Overture. Elle considère qu'une page est pertinente si son concepteur est prêt à payer le prix pour figurer dans les premières réponses du moteur.

Ce qui est curieux c'est de constater que quelle que soit la famille d'indicateurs retenue, il y a un invariant : l'indicateur de pertinence est l'affaire du moteur de recherche. Il existe dans tous les cas UN indicateur de pertinence utilisé par le moteur de recherche pour hiérarchiser ses résultats. Cet indicateur est volontairement gardé opaque pour éviter qu'il soit copié par les concurrents ou détourné au profit de webmestres soucieux d'améliorer le positionnement de leur site web dans le classement du moteur de recherche. Aujourd'hui cette

génération d'indicateurs de pertinence traverse une phase de rendement décroissant. Les moteurs de recherche doivent investir de plus en plus pour se démarquer de la concurrence et maintenir leur avantage concurrentiel. Ceci peut favoriser l'émergence de nouvelles approches.

C'est sans doute dans cette perspective qu'il faut comprendre l'innovation introduite par le moteur de recherche Chinois Openfind[41] dont l'interface anglaise est mise en ligne en Août 2002. Ce projet repose sur un diagnostic extrêmement lucide pour l'époque.

- « Pour une même requête, différentes personnes obtiennent le même classement des résultats quand bien même leurs attentes ou leurs préférences sont très éloignées.
- Le classement est fixé et les utilisateurs n'ont pas de moyen de choisir différentes façons de classer les résultats renvoyés par le moteur de recherche. »

Ce moteur prototype propose à l'internaute de préciser le poids qu'il souhaite attribuer à un certain nombre de critères associés à un profil. La pertinence au sens d'Openfind est alors définie par combinaison linéaire de ces divers indicateurs à partir des pondérations respectives que l'internaute a exprimées pour chacun de ces critères. Ainsi deux requêtes identiques peuvent donner lieux à des résultats complètement différents en fonction des priorités de chacun. Dans la version bêta, l'algorithme, appelé PolyRank, propose un classement par score, par taille et par date. Ce projet fut visionnaire à l'époque. Il préfigurait un moteur « boîte à outils ». D'ailleurs les concepteurs précisent, dans leur communication, « *La technologie PolyRank fournit des classements sophistiqués et optimisés pour les résultats d'une recherche d'information. Actuellement, Openfind laisse l'utilisateur choisir un classement par date ou par taille en plus du classement par défaut. D'autres moyens de classement*

[41] Le site web de ce moteur n'est plus disponible en ligne. L'interface peut être visualisée en interrogeant le moteur de recherche web.archive.org qui conserve une « mémoire du web »
http://web.archive.org/web/20030201083415/www.openfind.com/en.note.html

seront proposés dans le futur, permettant à l'internaute de contrôler les paramètres de classement[42] ».

En dépit de son avancée marketing, ce moteur disparaît du cercle étroit des moteurs prétendant au statut de moteur majeur et retrouve en 2003 une interface purement chinoise.

Ce type de projet ouvre la voie à une nouvelle génération d'indicateurs de pertinence « orientés utilisateurs » permettant de passer de l'absolu au relatif, de l'universel au contextuel. Openfind est donc un bon exemple d'initiative entrant dans la case I du *Tableau 7*.

La seconde révolution des moteurs de recherche concerne le passage d'une requête *topic related* a une requête multicritères.

2. Vers une requête multicritères

Depuis les débuts des moteurs de recherche, on a considéré qu'une requête était exprimée à travers un certain nombre de mots-clés définissant le sujet pour lequel on souhaite obtenir des informations. Lorsqu'un internaute effectue une recherche d'information, il exprime son besoin en précisant de façon plus ou moins fine le sujet de sa recherche. Cette étape de spécification du sujet est décisive car elle va orienter la réponse qui lui sera renvoyée par le moteur de recherche. Une question vague va générer des réponses elles-mêmes vagues. Aujourd'hui tous les moteurs de recherche fonctionnent sur ce principe qui connaît des limites intrinsèques liées à la difficulté pour l'internaute de générer une requête propre non ambiguë. Certains perfectionnements des moteurs de recherche mettent alors en œuvre des techniques comme l'expansion de requête ou la catégorisation a posteriori (Boutin et Al.- 2006). Aucune de ces méthodes ne remet en cause la question telle qu'elle a été formulée. Aujourd'hui tous les moteurs de recherche majeurs fonctionnent sur ce principe.

Nous considérons que l'expression du besoin pourrait être affinée par l'internaute à travers l'expression de dimensions complémentaires et orthogonales au sujet de la recherche. Ainsi une page web peut-elle être décrite par la tonalité de son discours, son degré de subjectivité, son niveau de

[42] http://web.archive.org/web/20030201083415/www.openfind.com/en.note.html

lisibilité, son niveau d'accessibilité, son niveau de centralité. Chacun de ces aspects contribue à apporter un éclairage particulier. Il devient alors possible pour l'internaute de spécifier les types de pages web qu'il souhaite privilégier. En permettant à l'internaute d'exprimer davantage ce qu'il souhaite, le moteur de recherche aura plus de facilité à le satisfaire. Quelques initiatives existent en la matière dans le domaine de la recherche d'information.

Yahoo! à travers le Mindset[43] a développé une interface, présentée *Figure 7,* dans laquelle l'internaute exprime sa requête de façon classique à travers des mots clés. Une fois les résultats obtenus, un curseur[44] apparaît à l'écran. Ce curseur est polarisé à gauche par *shopping* et à droite par *researching*. Yahoo! est capable de positionner chaque document sur un continuum commercial-recherche. Ce curseur est à la disposition de l'internaute qui peut choisir entre des documents plus ou moins commerciaux. Une fois que l'internaute a activé le curseur, la liste se réorganise pour répondre à ses besoins. Le Mindset de Yahoo ! est un exemple d'initiative entrant dans la case III du *Tableau 7*. En effet l'internaute exprime le poids respectif qu'il attribue au critère *Shopping* ou *researching.*

Figure 7: un exemple de moteur à curseur : le Mindset de Yahoo!

Dans le même esprit, le méta moteur Clush[45] dont l'interface est fournie ***Figure 8*** dispose d'un curseur permettant à l'internaute de choisir les résultats selon qu'ils contiennent plutôt du contenu ou des liens.

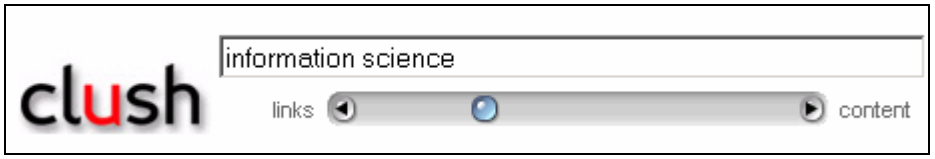


Figure 8 : un exemple de moteur à curseur : le métamoteur clush

[43] l'adresse url du mindset (mindset.yahoo.com) n'est plus active depuis 2008.
[44] On parle parfois de moteurs à curseurs pour désigner ce type d'initiative
[45] www.clush.com

Le moteur de recherche français Exalead[46] propose également une interface permettant dans sa partie droite d'affiner la recherche (voir Figure 9). Exalead, affiche plusieurs "visions" orthogonales des documents réponses à partir d'informations statiques (élaborées lors de l'indexation) ou dynamiques (élaborées à partir du contexte de la requête ou des réponses). Sous Exalead, cette approche par facettes est disponible dans les fonctions avancées du moteur. La raison essentielle de ce choix est la volonté de ne pas déstabiliser l'internaute par une interface trop riche. L'interface d'Exalead est un bon exemple d'initiative entrant dans la case II du Tableau 7.

Figure 9 : La fonction Affiner la recherche d'Exalead à partir de la requête « intelligence économique »

Le moteur de recherche blinkX TV[47] est un autre exemple d'initiative de ce genre dans le domaine de la recherche de vidéos. L'internaute a le choix entre accorder une importance plus grande au critère temps ou au critère pertinence. La *Figure 10* illustre l'interface présentée lorsqu'une requête est effectuée.

[46] http://www.exalead.fr
[47] http://emea-search.blinkx.com

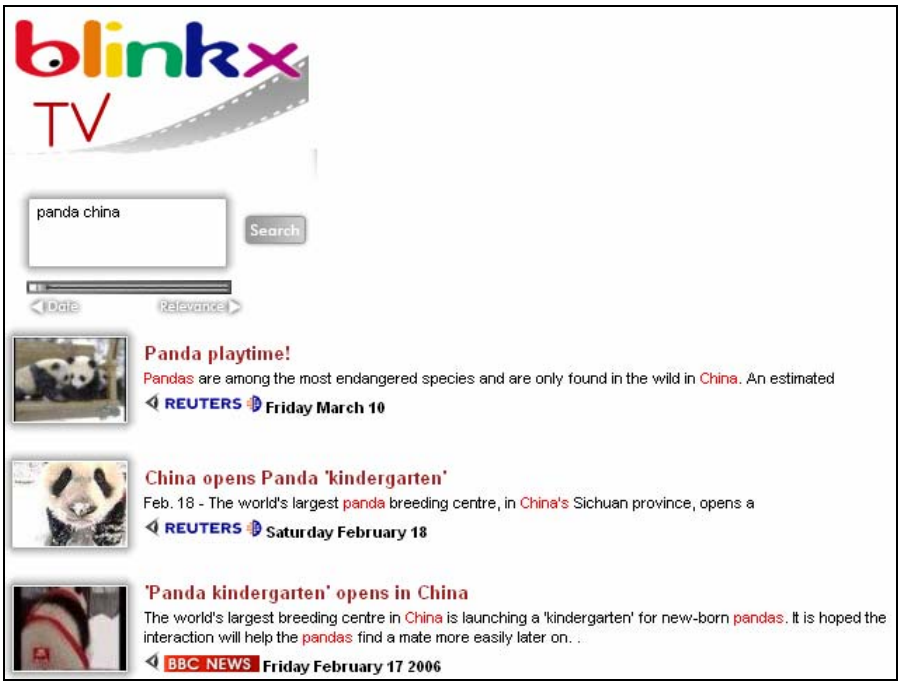


Figure 10 : un exemple de moteur à curseur : blinkx à partir de la requête « panda »

Dans le cas de like.com[48], qui est un outil de recherche d'objets graphiques, l'internaute dispose de trois curseurs (*color*, *shape*, *pattern*) pour faire varier l'indice de similarité des objets trouvés par rapport à la requête ou à l'objet initial. Cet outil applique des algorithmes de similarité à la recherche d'images. La *Figure 11* illustre le résultat renvoyé par cet outil.

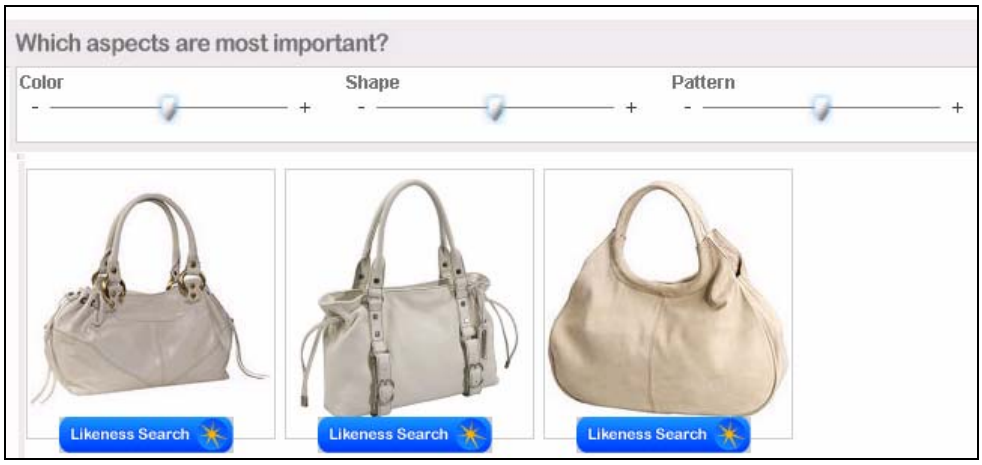


Figure 11 : un exemple de moteur à curseur : like.com

Ces mises en œuvre préfigurent ce que seront les moteurs à curseur, sans doute outils de recherche de demain. On peut penser dans l'avenir à un moteur qui,

[48] www.like.com

outre l'interrogation sous forme de mots clés, disposerait de plusieurs curseurs, chacun engageant l'internaute à se positionner en fonction de ses besoins : le résultat obtenu serait alors personnalisé. Le classement unique proposé par les moteurs de recherche traditionnels ne suffit plus. Deux personnes qui adressent la même requête au moteur de recherche (par exemple Dinosaure) obtiendront les mêmes résultats dans un moteur de recherche conventionnel alors que ces deux internautes peuvent être de profils très différents. Le premier est un écolier qui doit faire un exposé sur le sujet ; le second étudiant en paléontologie cherche des articles scientifiques. Aujourd'hui la recherche par mot clé est le seul moyen d'affiner le thème de la recherche pour la faire correspondre aux besoins.

Ces deux mutations des moteurs de recherche se renforcent mutuellement : en effet l'enrichissement de la requête par d'autres dimensions que celle du sujet de recherche est un processus qui conduit l'internaute à préciser ses besoins. Il va donc dans le sens d'un indicateur de pertinence centré sur l'utilisateur.
L'évolution tendancielle vers un rôle actif de l'internaute dans la construction de son indicateur de pertinence a une implication forte sur le fonctionnement des outils de recherche. Actuellement une énergie importante est consacrée par les moteurs de recherche à contrer les initiatives des référenceurs dont l'objectif est d'obtenir un positionnement idéal des sites de leurs clients sur les moteurs de recherche pour certains mots clés. Si chaque internaute construit son classement et si pour une requête donnée, il n'y a plus un mais des milliers de classements possibles (autant que l'on puisse obtenir par combinaison d'un nombre grandissant de curseurs), la question du référencement optimal devient insoluble.

Synthèse:
On observe aujourd'hui un décalage de culture au sein même du web entre deux grandes sphères toutes deux mobilisées par les mêmes internautes.
La sphère du « web 2.0 » sollicite l'internaute pour qu'il devienne acteur du système : les forums de discussion, les blogs, les logiques de *social tagging* s'inscrivent dans cette perspective. On parle par exemple de *Consumer Generated Media* pour désigner l'expression des clients sur le web. Un

sondage réalisé en 2006 par l'institut Ipsos[49] révèle que les Français accordent deux fois plus de crédit à l'avis d'autres consommateurs sur le net qu'à une publicité. De plus, 44% des internautes interrogés déclarent ne pas avoir acheté à cause de ce qu'ils ont pu lire sur Internet par d'autres clients ou particuliers.

Dans d'autres sphères du web, on observe des logiques dans lesquelles l'internaute est utilisateur et non acteur. C'est le cas aujourd'hui des moteurs de recherche grand public. Après avoir formalisé une requête adressée au moteur de recherche, l'internaute reçoit un ensemble de pages web hiérarchisées sans qu'il ait connaissance véritable des critères de classement ni qu'il puisse exprimer une quelconque préférence dans l'établissement de ce classement.

Le décalage observé entre ces deux sphères du web nous semble être le moteur d'une dynamique créatrice qui se traduit par la prise en compte d'un « souffle participatif » dans le domaine des moteurs de recherche. Cette logique s'exprime de deux façons complémentaires. D'une part certaines initiatives proposent à l'internaute de donner un poids à certains indicateurs permettant ainsi à l'internaute de se construire sa propre fonction de pertinence. D'autres initiatives proposent à l'internaute d'affiner sa recherche lorsqu'il effectue sa requête. Le moteur de recherche à facettes que nous proposons s'inscrit dans cette logique de moteur participatif. Nous serons conduits dans le chapitre 4 à préciser la notion de facette mais avant cela il est important de définir la méthodologie de ce travail de recherche.

[49] Quel est le pouvoir économique des Blogs en Europe institut Ipsos http://loiclemeur.com/IPSOSeuroblogs2006.pdf

CHAP 2 : METHODOLOGIE

Notre travail considère qu'une recherche d'information peut s'exprimer à travers plusieurs facettes orthogonales au sujet de recherche. Il a pour objectif de traduire sous forme d'indicateurs automatiques les dimensions souvent qualitatives permettant de caractériser une page web (niveau de subjectivité, valence, niveau de lisibilité, niveau d'accessibilité, niveau de centralité, niveau de vitalité d'une page web…).

Ce chapitre regroupe trois volets aux contenus bien distincts :

- Le premier a pour objectif de décrire la méthodologie expérimentale déployée au sein de ce travail de recherche pour évaluer la qualité de chaque facette retenue sur un échantillon test de pages web.
- Le second présente des pistes méthodologiques pour évaluer chaque facette dans la vraie vie, c'est-à-dire sur des pages web issues de recherches effectives.
- Le dernier volet constitutif de ce chapitre porte sur la présentation des méthodes que nous avons suivies pour constituer l'état de l'art ou plutôt les états de l'art car divers champs scientifiques ont été mobilisés.

A. Calibrage et évaluation de chaque facette à partir d'un échantillon d'experts

1. Problématique

La validation de la pertinence des indicateurs retenus peut s'effectuer de deux manières selon qu'il est possible ou non de comparer le résultat de l'indicateur au résultat d'une évaluation humaine : on effectuera alors une étude empirique ou analytique.

a) Evaluation analytique

Certaines facettes ne peuvent pas faire l'objet de validation par l'internaute (elles seront décrites plus en détail dans les chapitres suivants) :

- **Le niveau de fraîcheur d'une page web**. Cet indicateur correspond à la régularité avec laquelle cette page web est mise à jour.
- **Le niveau de centralité d'une page web** parmi d'autres pages web est utilisé en analyse de réseaux sociaux pour caractériser la position d'un point parmi un ensemble de points en interaction.
- **Le trafic d'une page web** est le nombre d'internautes ayant visité cette page durant une période donnée.

La mesure de la centralité d'une page web dans son contexte relationnel ne peut pas être définie par un internaute. Il est par ailleurs difficile de demander à un juge de reconstituer la centralité d'une page au risque de le faire travailler sur une base incomplète et subjective. Il en va de même de l'indicateur de vitalité d'une page web : l'internaute aura du mal à avoir une vision complète de l'évolution de la page car il ne porte qu'un regard instantané à l'instant t sur cette page. Enfin, le trafic d'une page ne peut être appréhendé par l'utilisateur sans indicateurs supplémentaires fournis par le serveur.

Dans ces différents exemples, l'indicateur obtenu automatiquement ne peut pas être confronté au jugement humain mais doit être évalué de manière analytique en se référant par exemple à d'autres indicateurs décrits dans la littérature.

b) Evaluation empirique

Dans d'autres cas, il est possible de juger de la concordance entre le résultat de l'indicateur automatique et le jugement exprimé par des évaluateurs humains. Voici quelques indicateurs qui seront décrits en détail au chapitre 4 :

- **La polarité d'une page web** (son orientation négative ou positive). Cette notion fera l'objet d'une analyse fine dans un chapitre ultérieur. Précisons à ce niveau que la polarité d'une page web correspond à la perception que va avoir l'internaute du caractère positif ou négatif de cette page. Plusieurs éléments entrent en ligne de compte pour construire cette perception comme le vocabulaire de la page ou le choix des couleurs. Dans une première vision très restrictive, nous proposerons de définir la polarité d'une page à partir de la valence des termes qui la constituent.

- **Le niveau de subjectivité d'une page web**. Nous aurons l'occasion de revenir en détail sur cet indicateur. Selon Wiebe (2000), l'objectivité se définit comme la capacité d'un document à présenter une information factuelle par opposition à une information subjective qui exprime une opinion ou une évaluation.
- **Le niveau d'accessibilité d'une page web**. L'accessibilité est définie par le W3C[50] comme la capacité d'un site web à être utilisé par une personne ayant une incapacité quelconque.
- **Le niveau de lisibilité d'une page web**. Selon Henry (1975), la lisibilité désigne « le degré de difficulté éprouvé par un lecteur essayant de comprendre un texte ».

Dans ce chapitre, supposons que nous disposons d'indicateurs quantitatifs. Nous proposons de décrire la **méthodologie générale** permettant :

- **Le calibrage des indicateurs** : en ajustant au mieux les résultats des indicateurs aux évaluations fournies par une base d'experts.
- **L'évaluation de la qualité des indicateurs** : en testant ces indicateurs de manière empirique à l'aide d'un échantillon d'utilisateurs.

Nous reprendrons plus en détail, chapitre 4, la méthodologie spécifique au traitement de chaque facette. Nous proposerons également des techniques de calcul de ces facettes.

2. Matériel et méthodes

Nous allons décrire ici la méthodologie expérimentale **générale** retenue :

a) Les sujets ou experts

Dans la partie expérimentale, nous recourons à des juges de langue maternelle française (nos experts). Nous ne pourrons donc pas valider les indicateurs pour d'autres langues. Toutefois, beaucoup de nos travaux sur les facettes sont issus

[50] World Wide Web Consortium

d'une littérature anglo-saxonne que nous avons cherché à transposer à la langue française[51].

L'expérimentation est réalisée à partir d'une population de 100 étudiants inscrits en seconde année de techniques de commercialisation de l'IUT de Toulon. Nos juges ne sont donc pas des professionnels de l'information et de la documentation. Toutefois, des expérimentations antérieures conduites dans le domaine de la mesure de la crédibilité d'une page web (Macedo Rouet, 2008) montrent que les réponses d'experts en information documentation vont dans le même sens que les réponses d'étudiants non experts. Les réponses des experts sont plus tranchées que les réponses des juges non experts.
Les étudiants sont répartis en 4 groupes de travaux dirigés. L'expérience se déroule au mois de janvier 2008. Cette expérimentation s'inscrit dans une démarche volontaire de l'étudiant. D'ailleurs tous les étudiants sollicités ont accepté de jouer le jeu. L'expérimentation se déroule pendant les 30 premières minutes de trois séances de travaux dirigés d'informatique de seconde année. L'intégration de cette expérience dans l'enseignement du cours d'informatique de seconde année se justifie par le fait qu'à travers ce travail, les étudiants sont conduits à se poser des questions sur les critères de validité de l'information web. L'objectif est de comparer des indicateurs automatiques de pages web au jugement humain. L'expérimentateur précise que les étudiants seront tenus informés des résultats de ce travail.

b) Matériel

L'expérimentation se déroule dans une salle informatique équipée de 15 ordinateurs connectés au web. Les étudiants sont assis par groupe de deux autour d'un ordinateur. Ce regroupement s'opère par affinité sans intervention de l'expérimentateur (ce travail par deux ne pose pas de problème particulier en ce sens qu'il s'agit au départ d'identifier un corpus de pages consensuelles). Chaque binôme d'étudiant se voit remettre une feuille comportant un tableau dans la première colonne duquel figure une liste de pages web à qualifier et en première ligne les différentes modalités qualitatives à choisir pour la ou les

[51] Par ailleurs, dans le cadre d'une collaboration internationale en cours avec L'école Normale Supérieure de Shanghai, nous travaillons à la transposition de ces indicateurs à la langue chinoise (Mandarin).

dimensions retenues. Le tableau mentionne la requête grâce à laquelle la page a été obtenue. Cela fournit donc un élément de contexte lors de la qualification de la page. Les étudiants ont également à leur disposition une version Excel du document papier qui leur est remis leur permettant de lancer la page sans avoir à en taper l'adresse. A l'issue du travail, ils doivent remettre la feuille qu'ils ont remplie et qui est anonyme. Un exemple de fiche remplie, dans laquelle on demandait aux juges de qualifier un corpus de 63 pages web selon le critère de valence et de niveau de subjectivité, est fournie en annexe 13. Dans la pratique, on a observé que tous les questionnaires ont été remis par les étudiants et que tous les questionnaires ont pu être exploités.

c) Consigne

Chaque binôme d'étudiants doit sélectionner 10 pages web de son choix parmi les pages web de la liste. Les pages choisies sont lancées sous le navigateur Internet Explorer. Le binôme attribue un jugement à la page en choisissant une modalité qui lui est proposée. Si les juges expriment la même modalité, celle-ci est cochée sur le tableau. En cas de désaccord inter juges, les deux modalités sont cochées. La signification des différentes modalités de la variable considérée est précisée oralement par l'expérimentateur aux étudiants par une définition type. Deux exemples polaires sont donnés pour illustrer les modalités opposées afin de bien faire comprendre l'indicateur qui est mesuré.

d) Choix des pages constituant l'échantillon.

Dans les différentes expérimentations que nous avons conduites, nous avons fait le choix de ne pas considérer un corpus de pages représentatif du web. Cette absence de représentativité peut apparaître comme un obstacle à l'inférence statistique c'est-à-dire à la généralisation au niveau de la population mère (les pages du web) d'observations qui se dégageraient de l'étude de ce corpus. En fait, ce choix permet d'augmenter la puissance statistique des tests utilisés pour évaluer la qualité de nos indicateurs. En effet, les indicateurs que nous déployons conduisent bien souvent dans la pratique à faire ressortir un petit nombre de pages web aux propriétés très particulières. Lorsqu'on s'intéresse, par exemple, à la polarité d'une page web, on observe typiquement 1% de pages très négatives. Il faudrait étudier plusieurs milliers de pages pour

espérer évaluer la qualité de cet indicateur si l'on voulait garder un échantillon représentatif. Or nous n'avons que 100 juges pour qualifier les pages web de nos corpus et plusieurs facettes à investiguer. Nos échantillons ont donc été construits pour révéler au mieux les facettes extrêmes des indicateurs sollicités.

e) Identification de pages web consensuelles

La qualité des indicateurs que nous avons construits s'apprécie par la concordance entre la valeur calculée par l'algorithme et le jugement d'usagers sur un ensemble de pages web (nos experts). Le problème est que tous les juges n'ont pas, par définition, la même appréciation d'une page pour le critère étudié et qu'on observe une variabilité interindividuelle. Pour cette raison, nous avons fait le choix de nous limiter à un corpus de pages consensuelles pour lesquelles les usagers expriment un choix voisin pour le critère considéré. Notre travail consiste donc dans un premier temps à définir le corpus de pages consensuelles puis à comparer dans un second temps l'avis des juges avec l'indicateur automatique[52]. Si on considère que l'écart entre les modalités de l'échelle est le même, alors la modalité exprimée par un juge peut être traduite par une note (entre 1 et 5 par exemple dans le cas où il y a 5 modalités). Si la même page est analysée par plusieurs juges, cette page sera donc décrite par plusieurs notes, chacune comprise entre 1 et 5. En utilisant un indicateur statistique de dispersion (écart type), on pourra juger du niveau de concordance entre les réponses des juges. Si l'écart type est nul, cela signifie que tous les juges ont exprimé le même jugement pour une page web. Plus l'écart type est élevé et moins les jugements sont consensuels.

f) Calibrage des indicateurs

A partir de cet échantillon de pages consensuelles est extrait de manière aléatoire un sous échantillon[53] utilisé pour **calibrer l'indicateur**. L'objectif est d'ajuster (ou étalonner) l'indicateur calculé sur les évaluations des experts. Les modalités d'évaluation étant qualitatives ordinales, nous proposons une

---

[52] Il serait intéressant d'observer comment réagit l'indicateur automatique dans le cas de pages non consensuelles. L'indicateur automatique prend il une position médiane ou une position plus tranchée ? Cette question ne sera pas étudiée dans ce travail

[53] La taille de cet échantillon dépend du nombre de modalités de la variable étudiée. Cicchetti et Fleiss (1977) proposent que la taille de l'échantillon soit supérieure à $2r^2$ où r désigne le nombre de modalités de la variable étudiée.

découpe de l'indicateur calculé (initialement quantitatif) afin de pouvoir obtenir les mêmes modalités. L'étape de calibrage consiste à trouver des valeurs de coupes "idéales" de l'indicateur de sorte à optimiser l'ajustement. Cette étape de calibrage est réalisée en cherchant à maximiser l'indice Kappa. Nous décrirons en détail l'utilisation de l'indice Kappa dans la suite du chapitre (mais également d'autres indices pouvant être utilisés).

g) Evaluation des indicateurs

Après avoir extrait un premier sous-échantillon de pages consensuelles pour calibrer l'indicateur, les pages restantes sont utilisées pour **évaluer l'indicateur**. L'indice Kappa (décrit paragraphe suivant) est à nouveau utilisé pour évaluer la concordance entre l'indicateur calculé de manière automatique (et rendu qualitatif dans l'étape de calibrage) et l'évaluation des experts. D'autres indices peuvent également être utilisés comme l'indice de rappel et de précision (décrits section suivante).

3. Indices et tests statistiques utilisés

a) Indice Kappa

La méthode que nous allons décrire ici a été utilisée non seulement pour calibrer chaque indicateur et trouver des coupes idéales (à partir d'un échantillon de pages consensuelles) mais également pour évaluer la qualité de chaque indicateur en se basant sur un nouvel échantillon.

Supposons que le calibrage ait déjà été effectué et que l'indicateur présente les mêmes modalités ordinales que l'évaluation des experts. Pour évaluer la concordance de ces deux variables nous proposons l'utilisation du test de concordance de Kappa. Le test non paramétrique Kappa (K) de Cohen (1960) permet de chiffrer l'accord entre deux ou plusieurs observateurs ou techniques (un observateur et une technique dans notre cas) lorsque les jugements sont qualitatifs. Cohen (1986) revient sur l'esprit de cet indicateur qui considère que l'accord observé entre des jugements humains et le jugement automatique résulte de la somme d'une composante « aléatoire » et d'une composante d'accord « véritable ».

Nous allons exposer la démarche de calcul de cet indicateur à partir d'une expérimentation pilote dont l'objectif était de s'intéresser à la concordance entre un indicateur de valence automatique (déjà étalonné) et l'appréciation par des juges humains (les experts) de la polarité d'un corpus de 29 pages web. Pour cet échantillon, nous disposons d'une classification des pages d'après l'indicateur automatique en trois catégories (positives, neutres, négatives) et d'une classification d'après les jugements humains en trois catégories également.

Le *Tableau 8* permet de confronter les jugements humains et automatiques. La concordance observée correspond au nombre d'individus qui sont classés dans la diagonale de la matrice (ici 19). Cette valeur rapportée au nombre de jugements totaux exprimés (ici 29) permet de connaître la proportion d'accords observés. Cette valeur notée (P_o) est de 65% dans notre exemple. Dans 65 % des cas, il y a concordance entre jugement automatique et jugement humain.

		Jugements humains sur pages consensuelles			
		Positif	Neutre	négatif	SOMME
Jugements automatiques	Positif	3	0	1	4
	Neutre	6	10	3	19
	Négatif	0	0	6	6
	SOMME	9	10	10	29

Tableau 8 Tableau de concordance entre jugement automatique et jugement humain : Présentation des valeurs observées sur un jeu de 29 pages web

L'accord aléatoire correspond aux valeurs espérées sous l'hypothèse d'indépendance des jugements. Cette valeur notée P_e est calculée dans le *Tableau 9*. L'effectif espéré pour une cellule est obtenu comme le produit des effectifs marginaux de la colonne et de la ligne correspondante divisé par l'effectif total (exemple $1{,}24 = 9 \times 4 / 29$) :

		Jugements humains sur pages consensuelles			
		Positif	Neutre	négatif	SOMME
Jugements automatiques	Positif	1,24	1,37	1,37	4
	Neutre	5,89	6,55	6,55	19
	Négatif	1,86	2,06	2,06	6
	SOMME	9	10	10	29

Tableau 9 : Valeurs espérées sous l'hypothèse d'indépendance des jugements : P_e

La concordance aléatoire notée P_e correspond au nombre de pages qui sont placées dans la diagonale de la matrice (ici 1,24 + 6,55 + 2,06 = 9,86) rapporté au nombre de pages au total (ici 29) soit $P_e = 0,34$.

L'indicateur de Kappa noté K est défini par la formule suivante :

$$K = \frac{P_o - P_e}{1 - P_e} \text{ soit } 0,46$$

Avec :

P_e =34% d'accord aléatoire

P_o= 65% d'accord observés

Le coefficient Kappa est un nombre compris entre -1 et 1. L'accord sera d'autant plus élevé que la valeur de Kappa est proche de 1 ; l'accord maximal est atteint (K = 1) lorsque Po = 1.

Landis et Koch (1977) ont proposé un classement de l'accord en fonction de la valeur de Kappa. Ce classement est proposé *Tableau 10*. Dans le cas de notre exemple (K=0,46), on observe un accord modéré.

Accord	Kappa
Excellent	≥ 0,81
Bon	0,80 – 0,61
Modéré	0,60 – 0,41
Médiocre	0,40 – 0,21
Mauvais	0,20 – 0,0
Très mauvais	< 0,0

Tableau 10 : Degré d'accord en fonction de la valeur de Kappa

Briand et al (2002) ont mis en évidence les biais associés à l'indicateur Kappa : « A proportions d'accords observés identiques, l'estimation du coefficient

Kappa est plus élevée lorsque les effectifs marginaux sont répartis de façon équilibrée sur les différentes modalités de réponse que lorsqu'ils présentent un fort déséquilibre en faveur d'une même modalité de réponse (déséquilibre symétrique) ». Or dans notre cas, les catégories ne sont pas équilibrées, la catégorie neutre étant survalorisée au détriment des catégories extrêmes.

Cicchetti et Fleiss (1977) proposent que la taille de l'échantillon de l'étude soit supérieure à $2r^2$ avec r le nombre de modalités de jugement. D'après Fermanian (1984), la taille minimale de l'échantillon devrait être 25 pour r = 3 et 30 pour r = 4 ou 5. Dans notre cas, on a un échantillon de 29 avec 3 modalités. Les conditions minimales de représentativité sont donc respectées.

b) Indices de « rappel » et de « précision »

Ces indices (Van Rijsbergen – 1979) sont complémentaires et couramment utilisés en « recherche d'information » pour qualifier la pertinence d'indicateurs ou pour mesurer la pertinence d'une classification. C'est essentiellement cette dernière utilisation qui va nous intéresser ici.

La « précision » se calcule comme le rapport entre le nombre de documents pertinents retrouvés et le nombre total de documents retrouvés. Le « rappel » correspond au rapport entre le nombre de documents pertinents retrouvés et le nombre total de documents pertinents. Considérons par exemple une classification (une de nos facettes) ayant pour objectif de distinguer des pages web ayant une caractéristique A de celles qui ne l'ont pas. Pour cela, un indicateur automatique mesure, pour chaque page, si on peut lui affecter le caractère A ou non. On demande en parallèle à des juges de préciser si la page a la caractéristique ou non. Les résultats peuvent être présentés Tableau 11 .

		Jugements humains	
		A	$\overline{A}$
Jugements automatiques	A	AA	$A\overline{A}$
	$\overline{A}$	$\overline{A}A$	$\overline{A}\overline{A}$

« Précision » : probabilité pour qu'un document retrouvé par l'indicateur soit pertinent. Une grande précision signifie que la plupart des items que la machine renvoie sont en adéquation avec le sentiment des juges.

$$\text{Précision} = \frac{AA}{AA + A\overline{A}}$$

« Rappel » : probabilité pour qu'un document pertinent soit retrouvé par l'indicateur. Un rappel élevé signifie que l'indicateur automatique a identifié la plupart des documents considérés comme pertinents par les juges.

$$\text{Rappel} = \frac{AA}{AA + \overline{A}A}$$

Tableau 11 : mise en évidence synthétique des concepts de « précision » et de « rappel »

Si on calcule cet indicateur pour la modalité « positive » de l'exemple présenté Tableau 8, on obtient une précision de 75% (3/4) et un rappel de 33% (3/9). Cela signifie que 3 pages web sur 4 sensées être positives le sont réellement mais que l'indicateur n'a identifié que 3 pages positives sur 9.

B. Confrontation de chaque facette à la réalité

Après avoir passé les étapes précédentes, nous souhaiterions confronter chaque facette à la vraie vie, c'est-à-dire à des pages pas forcément consensuelles, mais issues de recherches réelles. Nous allons identifier trois problématiques et décrire le protocole que nous proposons de suivre pour tenter d'y répondre.

1. Tester le pouvoir discriminant des indicateurs

L'objectif est alors de mesurer le pouvoir discriminant de l'indicateur. Cette étape peut être validée sans la présence de l'usager. Il s'agit de prendre un échantillon de pages représentatives du web, d'appliquer l'indicateur et de

dénombrer les pages pour lequel il a une valeur significative (dans un sens puis dans l'autre).

2. Evaluer l'efficacité de la recherche par facettes

Pour établir dans quelle mesure les nouveaux indicateurs élaborés sont de nature à créer un avantage concurrentiel durable pour le moteur de recherche qui les adopterait, il est nécessaire de mettre ce nouvel outil à l'épreuve de l'utilisation. L'objectif est alors de tester si le moteur à facettes permet de faire une recherche d'information plus efficace qu'un moteur de recherche classique.

L'efficacité de la recherche peut alors être définie par divers indicateurs :

- le temps de recherche avant de trouver une réponse satisfaisante.
- Le nombre de pages visitées avant de trouver une réponse satisfaisante

Pour apprécier ces éléments, 3 sources d'information peuvent être envisagées :

- Une étude quantitative reposant sur l'exploitation des fichiers log qui constituent une trace de la navigation de l'internaute sur le site. Pour exploiter cette source d'information, il faut comparer les traces laissées par les internautes sur le système à facettes avec les traces laissées par l'internaute sur un moteur de recherche traditionnel.
- Une étude qualitative s'appuyant sur l'exploitation de questionnaires administrés aux internautes ayant réalisé l'expérience pour apprécier leur niveau d'atteinte des résultats.
- L'analyse d'entretiens semi directifs permettant d'étudier la perception des internautes qui auront été exposés à un moteur à facettes.

Il est important de noter que la nouvelle interface proposée est assez différente de l'interface actuellement mise en œuvre par les moteurs de recherche. Pour cette raison, il serait intéressant de soumettre chaque internaute à plusieurs requêtes pour voir si un phénomène d'apprentissage a lieu et, si oui, au bout de combien de temps.

Pour mener à bien cet objectif, il faut disposer d'un prototype qui puisse permettre d'effectuer une recherche d'information dans des conditions se rapprochant le plus possible des conditions réelles. La réalisation d'un tel prototype suppose que soient résolus un certain nombre de problèmes :

- Il faut que cet outil s'inscrive dans une logique de temps réel. A défaut, il faut pouvoir pré-calculer les indicateurs pour un certain nombre de requêtes. Compte tenu de la chaîne de traitement semi automatique qui doit être déployée pour calculer les indicateurs correspondant à chacune des facettes, ce simulateur grandeur nature pour un nombre important de requêtes semble difficile à implémenter.
- Il faut que le prototype réalisé satisfasse à certains critères ergonomiques au niveau interface permettant de faciliter son appropriation par l'usager.
- Il faut concevoir un système permettant une traçabilité des usages.

Le niveau de finesse du prototype auquel nous sommes arrivés dans ce travail ne nous permet pas d'envisager de réaliser un comparatif de l'outil que nous avons déployé. Par contre, le prototype obtenu présente une vitrine de la technologie qui peut déboucher sur des demandes de financements ad hoc pour effectuer cette validation.

3. Evaluer l'adéquation des facettes à la vraie vie

Il s'agit d'une expérimentation qui a pour objet de définir les curseurs à retenir en fonction de la problématique de recherche. Le processus de définition des facettes a été élaboré en se confrontant à des univers de recherche d'information spécifiques (spécialistes de la veille, recherche d'information grand public) ou de travaux scientifiques particuliers. Quels échos ces facettes trouvent elles dans la réalité ? Un tel travail est nécessaire pour caler cette recherche dans la perspective d'un nouveau moteur de recherche généraliste ou plutôt une initiative qui va dans le sens d'une recherche d'information pour professionnels. Il faut donc concevoir une expérimentation pour justifier du choix des facettes ou l'adéquation des facettes avec des problématiques spécifiques.

Pour réaliser cette expérimentation, il faut pouvoir disposer d'un prototype élaboré. On pourrait alors observer le comportement des utilisateurs et voir quelle facette ils utilisent le plus volontiers dans leur recherche d'information. Comme dans le cas de l'expérimentation précédente, nous déportons la réalisation d'une telle expérimentation sur un travail ultérieur.

C. Méthodologie de construction de l'état de l'art

En théorie, tout travail de recherche scientifique s'appuie sur l'étude des travaux antérieurs. Cette étude de l'existant permet de produire l'état de l'art. Etape fondamentale d'une recherche, il permet de mettre en perspective les travaux antérieurs et de se positionner par rapport à eux. Toutefois rares sont les travaux de recherche qui définissent clairement la façon dont a été constitué le corpus qui a permis d'établir l'état de l'art[54]. Le caractère scientifique d'une recherche s'apprécie plutôt par la capacité du candidat à mobiliser pour son sujet une démarche scientifique qui va se retrouver au niveau de la question de recherche, des hypothèses, de la méthodologie déployée et de l'expérimentation. Un certain nombre de travaux en Sciences Humaines et Sociales définissent par exemple la démarche à suivre pour adopter une méthodologie quantitative ou qualitative, fournissant ainsi un cadre pour le chercheur. Un tel cadre ne nous semble pas exister pour ce qui est de la constitution d'une bibliographie reflétant l'état de l'art. Et pourtant des travaux dans le domaine de la scientométrie pourraient être utilisés et transposés au domaine qui nous intéresse de façon fructueuse.

Le processus de recherche débouchant sur la production scientifique comporte plusieurs étapes qui interagissent (*Figure 12*).

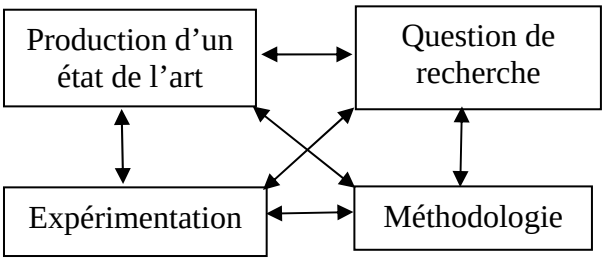


Figure 12: étapes dans la construction d'une recherche

Ces 4 étapes principales se nourrissent mutuellement. La qualité d'un document scientifique dépend de son maillon le plus faible. La fragilité de

[54] C'est un des objectifs des rapporteurs d'une thèse que de s'assurer que l'état de l'art recouvre correctement la diversité des travaux qui ont été conduits sur le sujet

l'état de l'art risque de compromettre la pertinence d'une recherche quand bien même celle-ci obéirait à une démarche scientifique sur son aspect méthodologique. Pour mettre en œuvre une méthodologie scientifique, il faut connaître les approches qualitatives et quantitatives et savoir les utiliser de façon pertinente dans le contexte de sa recherche. Pour faire un bon état de l'art, nous pensons de la même manière qu'il faut être guidé par une démarche. La démarche proposée par l'analyse scientométrique semble, de ce point de vue, pertinente. Or elle est peu utilisée.

Le chercheur constituant son état de l'art est souvent renvoyé à des pratiques empiriques que nous allons caractériser :

Parfois, en Sciences Humaines et Sociales, l'état de l'art est construit de manière intuitive sur un mode itératif à partir d'un premier corpus de références obtenues par une recherche bibliographique ou la consultation d'experts. Ce premier corpus est lu. Ces lectures permettent de dégager des mots clés nouveaux, de s'intéresser à d'autres publications du même auteur ou de sélectionner des références bibliographiques nouvelles qui vont permettre d'identifier d'autres documents qui enrichiront le corpus primitif. Intéressons-nous par exemple au processus par lequel le lecteur va accéder à d'autres ressources définies dans la bibliographie de l'article qu'il consulte. Le processus de sélection d'une des citations de la bibliographie d'un article dépendra du niveau d'attractivité du titre de la citation, du nombre de citations, de la fraîcheur de la citation, de l'intérêt du texte où figure le renvoi bibliographique, de la disponibilité du document sur internet, ou du budget du chercheur.

1. Inconvénients des pratiques empiriques de construction d'état de l'art

a) Résultats dépendants du corpus initial

Du fait de l'absence de méthode, le chercheur n'est jamais sûr d'avoir couvert les dimensions principales du problème. Les premières lectures en appellent d'autres selon une logique réticulaire qui peut prendre des dimensions exponentielles. Le chercheur est alors parfois conduit à privilégier un nombre restreint de documents. On pourrait comparer ce processus de constitution d'un

fond documentaire de référence à la mesure de la « saturation » calculée dans une étude qualitative : on continue d'interroger (de lire de nouveaux documents) jusqu'à ce que la personne interrogée marginale (le texte supplémentaire) apporte une quantité d'information supplémentaire faible par rapport à l'information collectée jusque là.

La qualité de la bibliographie dépendra donc in fine de la diversité et de la variété du fond documentaire primitif. Si, le chercheur réussit à identifier au départ un document de chacun des sous univers qu'il a à découvrir, alors il parviendra par propagation à collecter les articles au cœur du domaine. Si, au contraire, il se cantonne au départ à un petit nombre d'articles qui s'inscrivent tous dans la même logique, il aura du mal par la suite à retrouver dans son corpus final l'intégralité des sensibilités qu'un état de l'art devrait permettre de révéler.

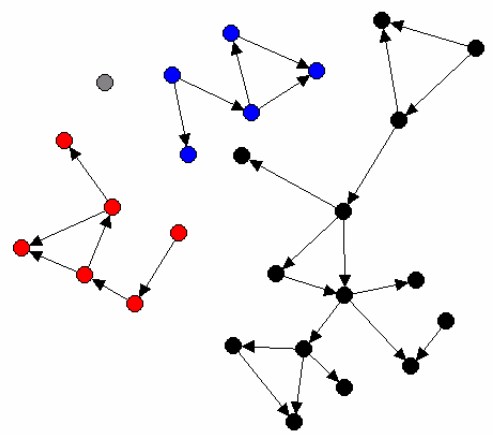


Figure 13: représentation des interactions entre documents constituant une bibliographie

La *Figure 13* fournit un processus imagé du type de fonctionnement que nous décrivons. Dans ce réseau, chaque sommet représente un article scientifique et chaque flèche représente la possibilité pour le lecteur de l'article de départ d'être orienté vers l'article d'arrivée. Cette orientation peut se faire par la bibliographie, par la découverte de mots clés dans un article qui va permettre d'élargir la recherche initiale à de nouveaux horizons, par la recherche d'autres articles du même auteur… Les démarches empiriques de constitution de l'état de l'art sont tributaires du ou des points d'entrée qu'aura le lecteur dans ce réseau. Dans notre exemple, le champ scientifique étudié est compartimenté en blocs qui ne communiquent pas entre eux. Dans un tel contexte, le chercheur doit avoir comme point d'entrée un sommet de chaque composante fortement

connexe du graphe (chaque sous graphe indépendant) pour espérer identifier tous ces nœuds. Si ce n'est pas le cas, il passera à côté de pans documentaires entiers de la littérature. On voit bien par ce graphe que du point de départ dépendront les points d'arrivée et que la reproductibilité du résultat n'est pas garantie.

b) Introduction de biais cognitifs

Le second problème, dont n'a pas forcément conscience le chercheur, est l'existence de biais cognitifs à l'œuvre dans sa recherche d'information. L'effet d'ordre (Asch, 1946) est un exemple qui s'applique particulièrement dans un contexte de surcharge informationnelle. Il correspond au fait que la première information qui est récupérée par le chercheur va avoir une place privilégiée dans son esprit. Le chercheur tente de retrouver, dans les documents suivants, une confirmation des idées originales contenues dans le premier document. Les idées nouvelles non confirmatoires risquent d'être écartées et non prises en considération. Difficile dans ces conditions de couvrir toute la variété d'un domaine.

c) Instabilité des résultats

Les deux problèmes évoqués ci-dessus conduisent à une absence de stabilité des résultats. Compte tenu du fait que chaque chercheur a une fonction de préférence spécifique, qu'il a un point d'entrée et des critères de navigation différents dans l'océan documentaire, il n'y a aucune certitude sur le fait que deux chercheurs effectuant la même recherche débouchent sur un état de l'art identique. Cette non reproductibilité de l'état de l'art est gênante d'un point de vue scientifique d'autant plus que l'état de l'art constitue le socle de la recherche sur lequel on s'appuie et par rapport auquel on se positionne. Cette absence de méthode oblige alors, pour compenser, à passer un temps disproportionné sur l'état de l'art qui se constitue par un processus de sédimentation s'étendant parfois sur plusieurs années.

Pour conclure sur ce point, on pourrait dire que le processus classique de constitution d'une bibliographie sur un sujet pose des problèmes de stabilité, de complétude et de biais par rapport au réel. Il laisse une grande place à

l'intuition, à la subjectivité, à la sérendipité. Cette approche humaine est nécessaire mais elle nous semble devoir être complétée voire précédée par une approche plus structurée qui mobilise le recours à des outils scientométriques.

Notre objectif est de préciser la façon dont nous envisageons la construction d'une bibliographie. Dans un premier temps nous définirons la méthode utilisée. Elle mobilise une démarche de collecte et de traitement de l'information. Cette méthode sera ensuite appliquée à la constitution de notre propre bibliographie.

2. Collecte de l'information

Il existe plusieurs sources « d'information secondaire » susceptibles d'orienter le chercheur dans la constitution de la bibliographie sur son sujet. Par source « d'information secondaire » nous entendons des dispositifs permettant à un chercheur de consulter une interface appropriée (une base de données bibliographique ou un moteur de recherche) pour obtenir des résultats intermédiaires (une notice bibliographique ou la liste renvoyée par le moteur de recherche) qui permettent à leur tour d'accéder directement ou non au document en texte intégral (qui constitue « l'information primaire »). Nous allons identifier trois de ces dispositifs pour étudier la place qu'ils peuvent occuper dans la constitution de la bibliographie sur un sujet : une base de données contenant des informations relatives aux citations, une base de données ne comportant pas d'informations relatives aux citations et un moteur de recherche généraliste. Comme nous allons le montrer, ces sources d'information secondaires sont partiellement redondantes : leur interrogation successive répond à un besoin de complétude sur le sujet car, comme on peut l'observer, leur focale est souvent étroite pour des raisons dues à des choix linguistiques ou de rang de revues.

a) Utilisation de sources comportant des citations

L'« analyse de la citation »[55] correspond à la valorisation rigoureuse et rationnelle de l'information contenue dans le champ références d'un corpus de références bibliographiques. L'analyse des citations induit toutefois plusieurs contraintes :

- La première est qu'il faut pour cela disposer de sources d'information particulières souvent coûteuses. Or peu de base de données fournissent des informations sur les références bibliographiques des notices bibliographiques. Nous sommes donc tributaires pour ce travail de bases de données souvent américaines du type Science Citation Index (SCI).
- Cette première restriction en induit une seconde. Les bases de données du Science Citation Index raisonnent sur les revues scientifiques de rang élevé. D'autre part, cette information renvoie à une littérature qui valorise surtout les publications anglo-saxonnes. Deux problèmes se posent donc. Le premier est qu'en privilégiant les articles d'excellence, on se coupe d'un nombre important de contributions scientifiques de rang inférieur. Le second est celui de la représentativité que peut offrir le SCI dans des disciplines comme les sciences de l'information et de la communication qui sont une spécificité française. Les publications des chercheurs français en SIC ne sont pas forcément représentées dans les bases du SCI. Se baser uniquement sur le SCI (pour des raisons scientométriques) pour construire son état de l'art conduit donc à choisir implicitement un ancrage et un positionnement par rapport à des travaux souvent anglo-saxons. C'est donc, dans notre discipline, sans doute se couper d'un ancrage national. Il serait donc judicieux de pouvoir compléter cette approche SCI par l'utilisation d'autres bases de données. Or il n'en existe pas de française qui comporte des informations de type citation.

[55] « L'analyse de la citation » étudie les références bibliographiques citées par l'auteur d'une publication scientifique

b) Recours à une base de données bibliographique française

Pour dépasser les limites de l'information disponible dans la base du SCI, nous complétons l'approche de la base du SCI par une approche complémentaire reposant sur des références bibliographiques issue de la base de donnés Pascal[56]. Cette base de données offre une couverture intéressante des revues de rang élevé couvertes par les bases du SCI mais prend également en compte des références d'articles en français ou dans des revues non indexées par le SCI. Il s'agit donc pour le chercheur français en SHS d'une information complémentaire à la base du SCI. Dans notre cas, une interrogation de la base Pascal nous a, entre autre, permis d'identifier un article en français de Maniez publié en 1999 consacré à la théorie des facettes.

c) Recours au web.

Le web est un outil de plus en plus sollicité par le chercheur pour obtenir des références bibliographiques ou des textes intégraux sur son sujet. Nous allons distinguer ces deux points.

La mise à disposition de notices bibliographiques est de plus en plus proposée par certains portails de grandes sociétés d'édition (Science Direct), certaines bases de données bibliographiques (ERIC), certains services web spécifiques (Citeseer), certains groupements de bibliothèques (SUDOC) ou certains sites marchands (Amazon.com). Ces sites web utilisent de plus en plus un langage partagé pour décrire ces ressources documentaires. Ce langage se présente sous forme de « méta données »[57] qui reprennent les champs des notices bibliographiques affichées à l'écran. Il existe maintenant en ligne un certain nombre d'outils gratuits qui permettent d'interroger ces sites web structurés autour de méta données bibliographiques et de récupérer ainsi des bases bibliographiques sur le domaine à explorer. Les informations étant homogènes, ces logiciels peuvent s'exécuter à partir de nombreuses bases de données permettant ainsi de se constituer une collection de références bibliographiques. Nous

[56] PASCAL est une Base de données bibliographiques multidisciplinaire et multilingue créée et mise à jour par l'Institut de l'information scientifique et technique.
[57] Une méta donnée est une donnée qui sert à décrire une autre donnée.

avons travaillé sur un de ces outils, Zotero[58]. En interrogeant des sites web adéquats, il est possible de collecter grâce à Zotero une bibliographie élargie qui servira de grille de lecture de documents primaires. Une telle recherche conduit à des résultats beaucoup plus volumineux que ceux que l'on obtient en explorant la base du SCI. En effet, les ressources web disponibles permettent de renvoyer à des articles qui ne font pas forcément tous parties des grandes revues qui seules sont indexées dans le SCI. Cette information permettra donc de récupérer des données structurées permettant de réaliser certaines des analyses automatiques (réseau des co-auteurs) et aussi d'éditer en fin de recherche une bibliographie au format de son choix (APA[59]…).

La mise à disposition des articles en texte intégral correspond à la logique individuelle du chercheur qui partage le texte intégral d'un article avec la communauté scientifique en déposant son article sur des Archives ouvertes (de type Hal[60] ou Archivesic[61]) ou sur un site web. Le web entre de plus en plus comme modalité de départ de constitution d'une bibliographie. Lawrence (2001) a montré que dans le domaine de l'informatique, la probabilité pour un article disponible en archive ouverte d'être cité est trois fois plus élevée que la probabilité pour un article disponible uniquement en format payant. Kurtz et al. (2003), Kurtz (2004) ont obtenu des résultats analogues dans le domaine de l'astrophysique et Odlyzko (2002) en mathématiques. Brody et al. (2004) a montré que le délai de citation de l'article est réduit de 3 ans lorsque le papier est déposé sur une archive ouverte.

[58] Zotéro est une extension du navigateur Firefox qui permet de récupérer et de gérer des références bibliographiques sur certains sites web. http://www.zotero.org/

[59] l'American Psychological Association a défini un ensemble de normes d'écriture des références bibliographiques appelé normes Apa. http://www.apa.org/

[60] HAL : archive ouverte permettant au chercheur de déposer sa publication au format de son choix (notice bibliographique, texte intégral) dans un espace consultable par d'autres chercheurs à travers une interface web. http://hal.archives-ouvertes.fr/

[61] Archivessic est une archive ouverte dans le domaine des sciences de l'information et de la communication archivesic.ccsd.cnrs.fr/

Synthèse :

Les trois sources « d'information secondaire » que nous avons présentées sont complémentaires comme l'illustre la Figure 14. La base du SCI permet d'identifier des acteurs majeurs de la communauté anglo-saxonne. Une base de données comme Pascal permettra de révéler une sensibilité francophone et de faire apparaître des travaux publiés dans des revues de rang inférieur. Le web sera utilisé pour rechercher de l'information primaire par une requête adressée au moteur de recherche. Il permettra aussi d'accéder à une information scientifique émergente qui n'a pas franchi toute la chaîne éditoriale d'une revue papier.

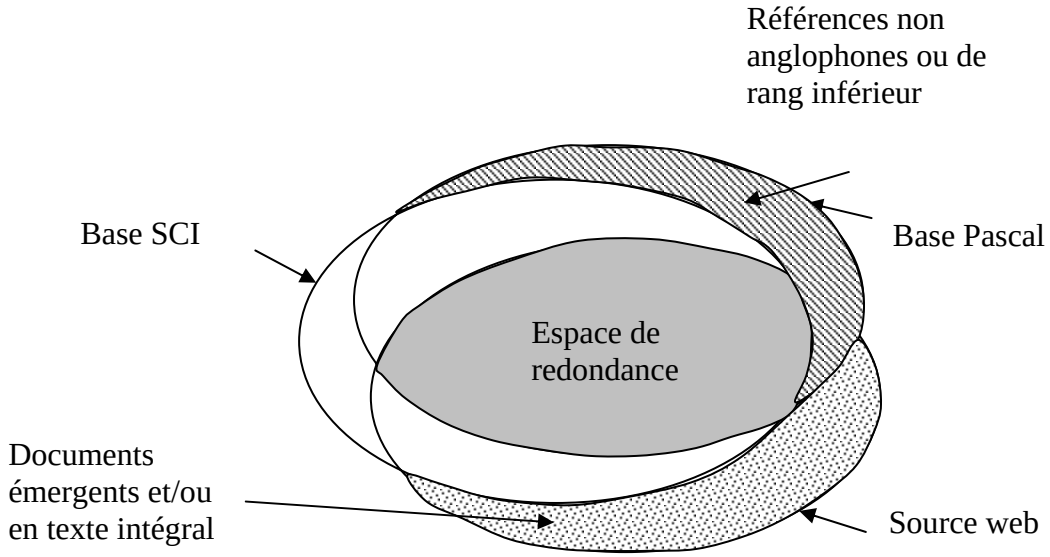


Figure 14 : Représentation figurée de la redondance entre source d'information et de l'espace spécifique de chaque source

3. Traitement de l'information

Nous mettons en œuvre deux grandes familles de traitement de l'information : d'une part l'analyse de la citation qui est déployée sur l'information collectée de la base SCI, d'autre part, l'analyse des réseaux de co-auteurs déployée à partir des données récupérées sur le web par Zotero.

a) Analyse de la citation.

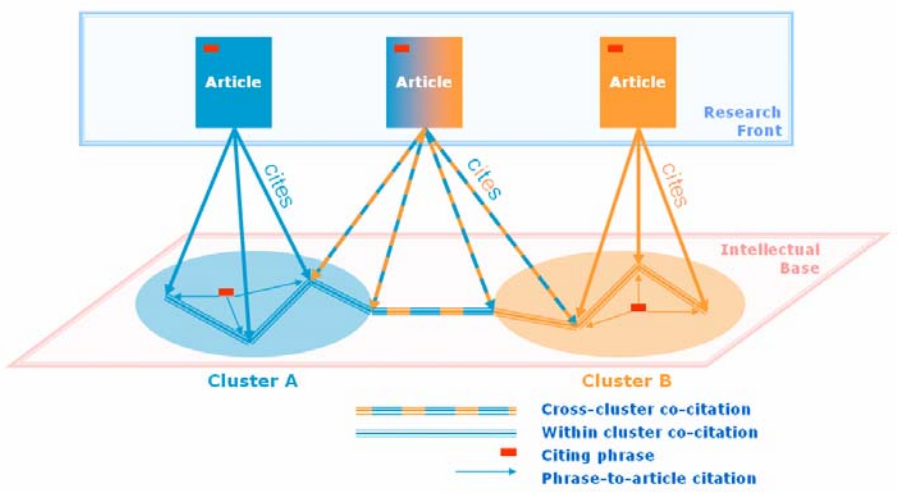


Figure 15: front de recherche et base intellectuelle

Le concept de « front de recherche » a été introduit par De Solla Price (1965) pour désigner un ensemble de documents émergents d'un champ de recherche. Pour Chen (2006), le front de recherche est l'état de l'art d'une spécialité à un instant t. Dans un domaine donné, le front de recherche est composé d'un corpus d'une cinquantaine d'articles récents. Persson (1996) distingue le front de recherche constitué des articles citants et la base intellectuelle (cœur de la littérature scientifique) composée des articles cités (*Figure 15)*. La base intellectuelle désigne pour Morris et al. (2003) un ensemble de documents de base fixes et invariables au cours du temps. Ce cœur est utile à tout nouveau développement. Pour Chen (2006), l'étude du front de recherche est l'analyse duale de la base intellectuelle. Ces empreintes peuvent être observées à travers l'analyse de type couplage bibliographique ou analyse des co-citations.

- Le couplage bibliographique (Kessler, 1963) s'intéresse au nombre de références que deux articles ont en commun. Ce type d'approche permet de structurer les documents constituant le front de recherche. Deux articles seront d'autant plus proches qu'ils auront un grand nombre de références en commun dans leur bibliographie.
- L'analyse des co-citations (Small, 1973) privilégie l'espace des références et relie deux références lorsque celles-ci figurent dans au moins deux documents faisant partie du front de recherche.

Cette articulation entre documents constituant le front de recherche et cœur de littérature peut être représentée de façon duale en privilégiant l'interaction entre documents du front de recherche sur la base du nombre de citations qu'ils ont en commun ou alors l'interaction entre documents constituant le cœur de la littérature sur la base du nombre de fois où ils sont cocités dans le front de recherche. La *Figure 15* fournit une carte des interactions entre les documents répartis entre « front de recherche » et « base intellectuelle ». Dans la réalité, le chercheur n'a pas cette carte en main et navigue à vue. Il nous semble intéressant de permettre au chercheur de se construire, sur son sujet, la carte des interactions entre les documents de son corpus.

b) Construction du réseau des co-auteurs

Dans ce cas, il n'est pas nécessaire de disposer d'information sur les citations. Les documents sont rapprochés sur la base d'auteurs qu'ils ont en commun. L'information peut être représentée sous forme de réseau.

On pourrait également, selon un principe voisin, mesurer les relations entre articles en évaluant directement la ressemblance des idées abordées dans les textes. Cette méthode repose sur l'exploitation des champs mots-clés ou codes de classification. Ces descripteurs sont à présent considérés comme les meilleurs représentants des concepts abordés dans les articles.

Cette approche relationnelle a deux finalités principales.

- Il peut s'agir d'identifier au sein d'un corpus documentaire des sous corpus denses qui correspondent à des sensibilités différentes. Ceci permettra d'appréhender dès le départ la variété des sujets couverts.
- Pour chaque grand thème ainsi identifié, il convient de faire ressortir des sommets aux propriétés particulières. Il peut s'agir de travaux qui se situent au cœur de la problématique pour orienter les lectures futures vers ces documents. Des travaux charnières entre plusieurs domaines peuvent être également extraits à travers la notion de point d'articulation du graphe.

4. Validation expérimentale

Le sujet de recherche qui nous préoccupe se caractérise par son caractère multidisciplinaire. La bibliographie de ce travail de recherche est associée à différentes micro bibliographies. Dans cet exemple et pour illustrer la démarche, nous considérerons la bibliographie générale relative à la théorie des facettes. Voici le protocole utilisé :

Le 3 Octobre 2007, nous récupérons le résultat d'une interrogation du *web Of Science*. La requête est la suivante : facet classification ; elle renvoie 181 références bibliographiques.

Dans la *Figure 16*, nous représentons, avec le logiciel NetDraw, le nombre de références bibliographiques que partagent les articles deux à deux. Chaque sommet du graphe représente le numéro d'un document qui traite du sujet. Un arc entre deux sommets signifie que les deux papiers ont au moins une référence en commun. Examinons la topographie de ce réseau et conduisons un raisonnement par l'absurde. Si la notion de « *facet classification* » correspondait à une expression univoque stabilisée dans la littérature, alors les documents traitant de ce sujet feraient référence à un petit nombre d'auteurs souvent identiques. Cela se traduirait par un enchevêtrement très dense du réseau représenté *Figure 16*. Or le réseau obtenu est plutôt lisible et se compose d'une composante fortement connexe au centre, de sommets isolés à gauche et de sommets peu liés entre eux au milieu. Le fait que « facet » appartienne au langage courant explique la présence de ces articles isolés qui n'ont aucune référence avec les autres. Ces articles peuvent être exclus de l'analyse.

Si on étudie la structure de la composante fortement connexe principale, on observe que celle-ci est organisée autour de sous groupes particulièrement denses. Il semblerait qu'au sein de chaque sous groupe, on parle d'un même sujet puisqu'on fait référence aux mêmes articles en bibliographie. Ce réseau a servi de grille de lecture de document primaire. Au lieu de consulter chacune des notices figurant dans ce réseau, on s'est concentré sur une ou deux références bibliographiques de chaque sous groupe. Cela nous a permis assez

rapidement d'identifier un sous groupe de 17 documents qui correspondent au sujet.

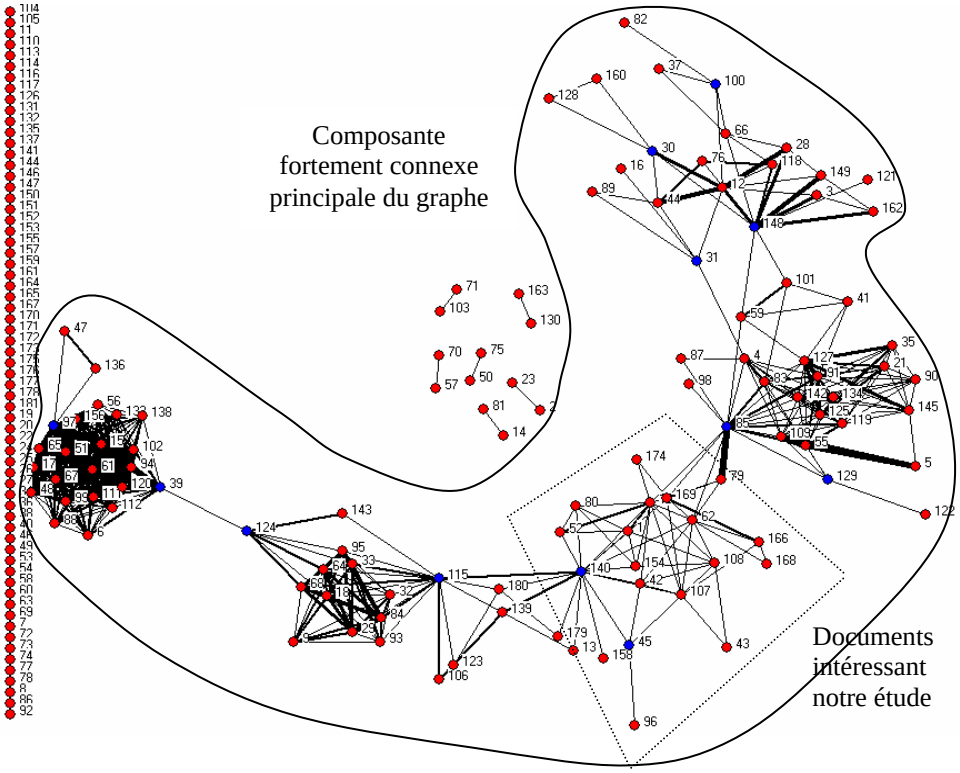


Figure 16 : Dessin du couplage bibliographique par le logiciel netdraw

Ces 17 articles constituent le front de recherche. Les interactions entre ces articles sont fortes et ne laissent pas apparaître de sous groupes disjoints. Ces articles sont représentés à l'intérieur d'un polygone (voir *Figure 16*).

Nous nous sommes intéressés *Figure 17* aux pays d'appartenance des auteurs de ces articles. On s'aperçoit que la théorie des facettes est développée dans 10 pays différents.

Figure 17 : pays d'appartenance des travaux caractérisant le front de recherche

On peut de la même manière s'intéresser *Figure 18* à l'année de publication des recherches constituant le front de recherche. Cela permet de juger de son ancienneté.

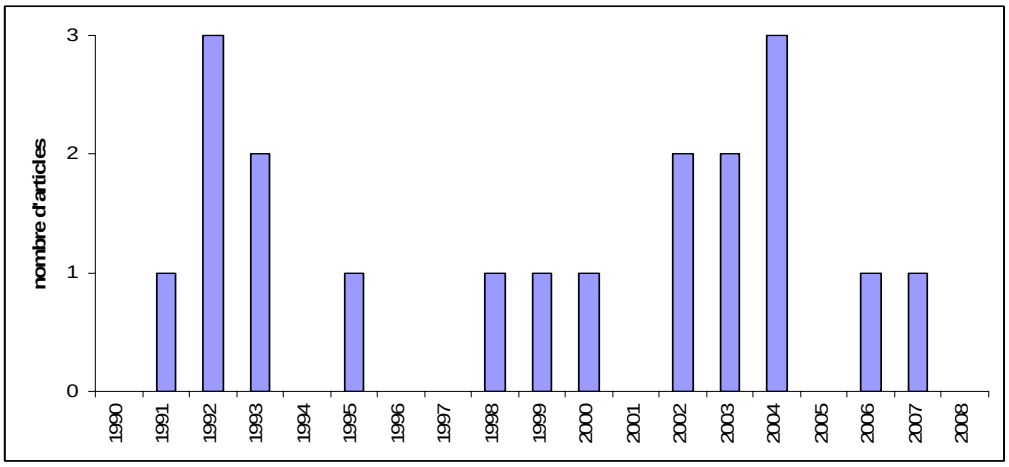


Figure 18 date de dépôt des travaux caractérisant le front de recherche

On a donc une dispersion dans le temps des dépôts d'articles sur cette thématique avec deux pics en 1992 et 2004. Nous pouvons également privilégier (*Figure 19*) l'étude des revues scientifiques dans lesquelles ces travaux ont paru.

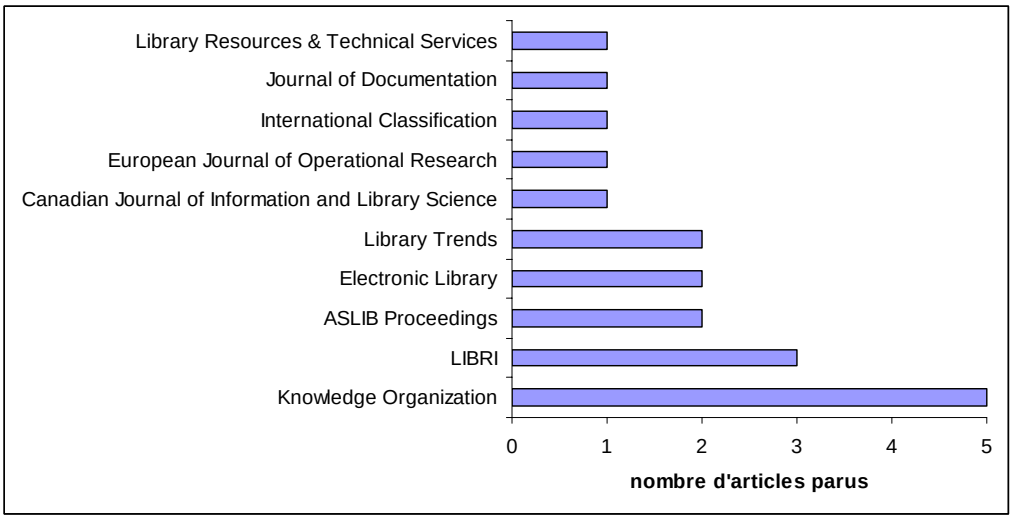


Figure 19: nom des revues où sont déposés les documents constituant le front de recherche

Ces statistiques permettent d'identifier les revues pertinentes pour le sujet et d'effectuer une veille sur ces revues. A l'issue de cette étape, nous sommes en mesure d'identifier des références bibliographiques qu'il faut intégrer dans nos lectures ; Elles figurent annexe 1 (pour ce qui concerne le front de recherche) et annexe 2 (pour ce qui concerne les articles constituant la base intellectuelle).

Nous allons maintenant nous intéresser à l'analyse duale de la première. Elle consiste à travailler sur l'espace des citations. Les 12 auteurs qui sont cités dans plus de 4 références (*Tableau 12)* peuvent être considérés comme les pères fondateurs du domaine.

Auteurs	Citations
RANGANATHAN SR	11
VICKERY BC	8
HJORLAND B	5
ELLIS D	5
DAHLBERG I	4
SATIJA MP	4
FOSKETT AC	4
AUSTIN D	4
MILLS J	4
FOSKETT DJ	4
GOPINATH MA	4
INGWERSEN P	4

Tableau 12 : auteurs cités dans plus de 4 références

Dans le *Tableau 13*, nous passons d'une analyse au niveau des auteurs à une analyse plus fine au niveau de leurs publications. Les auteurs les plus cités sont définis avec un fond gris dans le tableau ci-dessous. Ce tableau permet de voir quels sont les articles les plus présents dans les références. Il faudrait récupérer les références de ces auteurs pour compléter nos lectures (cases grisées). Il faudrait aussi considérer les articles des auteurs Robets et Stiles car on a là des articles de fond qui ont été cités un grand nombre de fois par la communauté. Certains auteurs du *Tableau 12* (Dahlberg, Foskett et Gopinath) ont eu une forte production scientifique et les références ne se rapportent jamais au même document. Nous ne les avons donc pas retenus.

Publications	Citations
RANGANATHAN SR, 1960, COLON CLASSIFICATION	5
INGWERSEN P, 1992, LIBRI, V42, P184	4
VICKERY BC, 1960, FACETED CLASSIFICATI	4
ROBETS N, 1969, J LIBR, V1, P131	3
RANGANATHAN SR, 1937, PROLEGOMENA LIB CLAS	3
STILES WG, 1985, CAN J INFORM SCI, V10, P16	3
FOSKETT AC, 1996, SUBJECT APPROACH INF	3
ELLIS D, 1999, ASLIB PROC, V51, P3	3
MCILWAINE IC, 1999, KNOWL ORGAN, V26, P23	2
MILLS J, 1977, BLISS BIBLIO CLASSIF	2
LIU SQ, 1990, INT CLASSIF, V17, P14	2
DAVIES R, 1989, J DOC, V45, P273	2
GODERT W, 1991, INT CLASSIF, V18, P98	2
NEELAMEGHAN A, 1991, INT CLASSIF, V18, P92	2
AITCHISON J, 1986, J DOC, V42, P160	2
BHATTACHARYYA G, 1979, LIBRARY SCI SLANT DO, V16, P1	2
SATIJA MP, 1992, INT CLASSIF, V19, P3	2
RANGANATHAN SR, 1933, COLON CLASSIFICATION	2
COSGROVE SJ, 1992, LIBR HI TECH, V10, P33	2
WATTERS CR, 1987, INFORM PROCESS MANAG, V23, P433	2
GODERT W, 1987, Z BIBLIO BIBLIO, V34, P185	2
ELLIS D, 2000, J INTERNET CATALOGIN, V2, P97	2
HJORLAND B, 1998, KNOWL ORGAN, V25, P162	2
STRAUSS A, 1994, HDB QUALITATIVE RES	2
KELLY GA, 1955, PSYCHOL PERSONAL CON	2
*CLASS RES GROUP, 1955, LIB ASS RECORD, V57, P262	2
AUSTIN D, 1984, PRECIS MANUAL CONCEP	2
RANGANATHAN SR, 1967, RANGANATHAN SERIES L, V20	2
SVENONIUS E, 1992, LIBRI, V42, P176	2
VICKERY A, 1987, INFORM PROCESS MANAG, V23, P99	2

Publications	Citations
VICKERY BC, 1958, CLASSIFICATION INDEX	2
DEVADASON FJ, 1985, INFORM PROCESS MANAG, V21, P11	2
DEWEY M, 1996, DEWEY DECIMAL CLASSI	2
RANGANATHAN SR, 1963, 5 LAWS LIBRARY SCI	2

Tableau 13 : articles des pères fondateurs

A l'issue de cette étape, nous sommes en mesure d'identifier des références bibliographiques qu'il faut intégrer dans nos lectures en Annexe 2 :

Synthèse :

Ce premier travail permet donc d'identifier un certain nombre d'articles de référence. On s'aperçoit qu'ils appartiennent tous à un petit nombre de journaux majeurs dans la communauté. La collecte de documents a été complétée par une recherche effectuée à partir d'autres bases de données (Pascal) et de moteurs de recherche web. Ces interrogations complémentaires ont permis d'identifier un corpus de documents de départ. La lecture de ces documents invite, en fonction de leur intérêt respectif, à approfondir des pistes en suivant, selon le cas, des liens en bibliographie, en recherchant d'autres articles du même auteur, en suivant la production des co-auteurs d'un auteur intéressant ou en réinjectant des termes spécifiques propres au jargon d'un chercheur.

CHAP 3 : ETAT DE L'ART

Ce travail de recherche fait référence aux notions de « facette » et de « recherche d'information ». Il convient donc désormais de préciser successivement ces deux notions pour voir de quelle façon nous envisageons d'étudier la recherche d'information au prisme de la théorie des facettes.

A. La notion de facettes en recherche d'information

La notion de facettes en Science de l'information a été développée, au départ, dans les travaux de Ranganathan (1937). Lorsqu'on étudie la littérature de Ranganathan, on constate, comme a pu le faire Spiteri (1998), que ses textes sont difficiles à appréhender. Cela se traduit par une certaine ambigüité dans l'usage qui est fait du terme facette. Cette ambiguïté trouve sans doute son fondement dans la volonté de Ranganathan de proposer une version universelle de la théorie des facettes. Elle est accentuée par le fait que le terme de facette appartient au langage courant. Il est important de commencer par clarifier le concept de classification par facettes. Nous proposons de le faire de deux manières :

- Dans un premier temps, nous positionnerons la classification par facettes par rapport à d'autres systèmes de classification. Cette approche débouchera sur une perception intuitive de ce système de classification et de ses avantages par rapport à d'autres techniques de classification.
- Dans un second temps, nous définirons la classification par facettes de l'intérieur par les caractéristiques qui la fondent. Cette approche permettra de comprendre les ambiguïtés du terme. A l'issue de ce chapitre, nous proposerons une définition de la théorie des facettes qui sera retenue dans la suite de ce travail.

1. Classification

L'objectif de ce paragraphe est de décrire ce qu'est une classification, comment elle est construite, quelle est son utilité. La science de la classification a pour objectif de définir les éléments élémentaires univoques de la connaissance et les relations qui existent entre ces éléments de connaissance. C'est le sens des propos de Farradane (1957) qui précise qu'« un tout est davantage que la somme des parties. C'est pour cette raison qu'il est nécessaire de considérer les relations entre concepts élémentaires pour lier les parties dans un tout cohérent ». La science de la classification a connu des développements différents selon le domaine qui fait l'objet de la classification.

- Dans le domaine des sciences naturelles, il s'agit de mettre des éléments clairement identifiables dans des catégories.
- Lorsque la classification s'applique aux idées, elle dépend davantage de l'interprétation ou des jugements.

Garfield (1957) distingue les classifications a priori des classifications a posteriori. Dans le cas des classifications a priori, la construction des classes est préexistante. L'analyse de la citation est un exemple de classification a posteriori. Cette analyse peut permettre a posteriori de regrouper des documents qui partagent des citations voisines. Les classifications a posteriori sont plus susceptibles de générer des associations émergentes ou innovantes. Pour Vickery (1957), l'analyse de la citation donne les signes avant coureurs de l'interaction entre des domaines jusque là séparés.

Mai (2004 b) distingue deux approches théoriques de la classification.

- La première est déductive. Les objets sont regroupés dans des classes sur la base de leurs propriétés observables. Pour Mai (2004a), la structure de la classification doit refléter un ordre préexistant défini a priori. Une classification erronée sera rejetée dès qu'une exception sera identifiée. Ainsi la classification représente, à terme, la structure de la connaissance. Elle a pour objectif de développer un système qui reflète la réalité (Feinberg-2007). La classification a pour objectif d'encapsuler

la variété des perspectives d'un domaine particulier de manière aussi fidèle que possible.

- La seconde considère la classification comme une vue sur le monde, une interprétation du monde (Mai -2004a). Cette interprétation du monde dépend de facteurs culturels. Couzinet (2006) en fournit quelques exemples. La classification russe BBK (bibliotetchno bibliographitcheskaya classifikatsiya[62]) propose plusieurs rubriques consacrées au marxisme léninisme là où les classifications américaines ne consacrent qu'une ligne à ce mot clé. Otlet et Henri Lafontaine, deux pacifistes du début du vingtième siècle qui ont fondé la CDU considèrent que l'établissement d'une classification internationale est un facteur qui devrait favoriser l'interdépendance, la communication intellectuelle et la solidarité (Otlet – 1934). La classification dépend aussi de la perception fine du domaine qu'aura la personne qui procède à la classification. Vickery (1957) précise que la manière de regrouper va dépendre de l'objectif visé par la personne qui classifie. Les types de bois seront regroupés différemment par un botaniste et par un architecte. Un domaine de la connaissance peut être classé selon différentes perspectives épistémologiques. Hjørland (1998) prend l'exemple de la psychologie et précise qu'une classification dans ce domaine travaille sur des concepts et non des éléments observables comme en sciences naturelles. Holmstrom (1957) précise que la classification de faits physiques peut être objective mais que la classification de documents relatifs à des idées ne peut pas être indépendante de ceux qui utilisent ces documents. Il n'y a donc pas une bonne classification qui serait vraie et les autres fausses. Le créateur d'une classification impose une vue particulière de la connaissance. La façon de catégoriser dépend de notre façon de voir le monde (Lakoff-1987). Shirky (2005) est beaucoup plus terre à terre. Il constate que les classifications sont souvent adossées à une bibliothèque (comme c'est le cas de la LCC adossée à la librairie du Congrès). L'objectif de la classification est alors d'optimiser le nombre de livres sur l'étagère. Les

[62] Littéralement « classification bibliothéco-bibliographique »

catalogues de bibliothèques ne sont pas là pour classer les concepts mais pour ranger les livres. Le schéma de catégorisation est alors avant tout une réponse aux contraintes physiques de stockage. L'auteur donne l'exemple du système de classification Dewey dans lequel la catégorisation des religions du monde est présentée dans le *Tableau 14.*

Dewey, 200: Religion
210 Natural theology
220 Bible
230 Christian theology
240 Christian moral & devotional theology
250 Christian orders & local church
260 Christian social theology
270 Christian church history
280 Christian sects & denominations
290 Other religions

Tableau 14 : les religions du monde dans la classification Dewey

A l'époque où la classification a été construite, il y avait sans doute autant d'ouvrages sur les « autres religions » que de livres sur la « théologie chrétienne ».

Le *Tableau 15* illustre la catégorisation de l'histoire dans la classification de la Librairie du Congrès américain.

DA: Great Britain	DK: Former Soviet Union
DB: Austria	DL: Scandinavia
DC: France	DP: Iberian Peninsula
DD: Germany	DQ: Switzerland
DE: Mediterranean	**DR: Balkan Peninsula**
DF: Greece	**DS: Asia**
DG: Italy	**DT: Africa**
DH: Low Countries	DU: Oceania
DJ: Netherlands	DX: Gypsies

Tableau 15 : Catégorisation de l'histoire dans la Librairie du Congrès

2. Les techniques de classification

Commençons donc par définir la « classification par facettes » par ce qu'elle n'est pas en la positionnant par rapport à d'autres logiques classificatoires. Il existe différentes techniques. Les techniques de classifications hiérarchiques sont les plus utilisées. La Dewey Decimal Classification (DDC) ou la Library of Congress Classification (LCC) sont des exemples majeurs de systèmes de classification hiérarchiques. Selon cette logique, décrite par Glassel (1998), construire une classification, consiste à lister tous les sujets possibles et à leur affecter un numéro d'ordre dans une classe. La connaissance est compartimentée

ex nihilo dans des catégories encapsulées dans des structures hiérarchiques. Par exemple, le code de la DDC pour désigner les maladies virales du riz est

633 189 8

633 = céréales

633.18 = riz

633.189 = maladies du riz

633.189.8 = maladies du riz d'origine virale

Utiliser une classification hiérarchique consiste à retrouver le « trou de pigeon » dans lequel se trouve la connaissance selon l'expression de Ranganathan en 1924 alors qu'il était étudiant.

La classification par facettes obéit à une autre logique. Il s'agit d'un langage documentaire qui permet de caractériser le document à décrire non plus par une seule dimension mais par plusieurs dimensions complémentaires. Chaque dimension appelée facette peut prendre différentes modalités. Ainsi un document sera décrit par la combinaison de plusieurs facettes.
On peut reprendre l'exemple des maladies virales du riz et voir de quelle façon ce code est traduit dans la classification Colon. L'indice le plus proche est :

EJ,381;421:5:

EJ = agriculture : facette principale

381 = riz : facette Personnalité

421 = maladie virale : facette Matière

5 = éradication : facette Énergie

Cette classification est dans son principe très ouverte puisque le champ des possibles qui peuvent être décrits est le produit du nombre de modalités de chaque facette. En cas de besoin, de nouvelles facettes peuvent être créées sans remettre en cause l'équilibre général de la construction.

Deux critiques principales peuvent être adressées aux techniques de classification hiérarchiques : leur caractère figé et la difficulté pour l'usager de rentrer dans la logique constitutive de la classification.

- Les classifications hiérarchiques correspondent à une vision du monde encyclopédique assez figée ou tout est prévu au départ. Les modifications de ces classifications sont difficiles car elles obligent parfois à sortir des grands principes que la classification avait érigés.

Elles se font parfois en rajoutant des rubriques à la marge mais ces ajouts successifs détériorent petit à petit la logique d'ensemble. Pour Glassel (1998), cette logique d'affectation a priori dans une classe est peu compatible avec l'expansion de la connaissance qui se traduit par la découverte de nouvelles idées, de nouvelles connaissances, de nouvelles façons de combiner ces connaissances nouvelles entre elles. De ce point de vue, la classification par facettes correspond à une vision du monde très ouverte dans laquelle il n'est pas nécessaire que les possibles soient prédéfinis à l'avance. Cette approche semble donc beaucoup plus souple pour permettre de décrire des contextes fortement évolutifs.

- Les schémas de classification traditionnels sont définis du haut vers le bas. Selon Shirly (2005), utiliser un système de classification traditionnel oblige l'utilisateur à abandonner sa vue sur le monde au profit de celle du système de classification. Shirly rapporte des résultats d'études qui montrent qu'il y a une concordance pauvre entre le classement d'un concept par l'utilisateur et son positionnement dans la classification. Si les classifications sont utilisées dans un contexte professionnel, alors l'utilisateur aura le temps de rentrer dans la logique de classification. Il en va autrement lorsque le système est utilisé ponctuellement par un novice : n'ayant pas investi le temps nécessaire, il aura du mal à retrouver ce qu'il cherche. Un système à facettes est de ce point de vue beaucoup plus adapté à un utilisateur non professionnel. En effet, un système à facettes permet de trouver une même information à partir de plusieurs chemins. L'existence de divers chemins facilite l'orientation de l'usager dans le système à facettes. Considérons par exemple un ensemble de recettes. On peut vouloir chercher une recette à partir de ses ingrédients, de la façon dont elle est préparée (grillée, en sauce, marinée), de l'occasion (Noël…), du pays où elle est pratiquée.

Pour avancer, il nous faut maintenant décrire plus précisément la notion de facettes. C'est ici qu'on retrouve une certaine ambiguïté dans les travaux scientifiques.

Denton (2003) distingue la classification hiérarchique unidimensionnelle de la classification par facettes qui est multidimensionnelle. La classification par facettes est adaptée au cas où il est possible d'organiser les entités en au moins trois catégories mutuellement exclusives et exhaustives dans la caractérisation du phénomène qu'elles cherchent à décrire.

D'après Bates (2002), la classification par facettes est à la classification hiérarchique ce que les bases de données relationnelles sont aux bases de données hiérarchiques."

3. La classification par facettes

Nous envisagerons successivement les origines de la théorie des facettes telles qu'elle a été définie par Ranganathan avant de nous intéresser aux perfectionnements ultérieurs

a) Origines de la théorie des facettes

S.R. Ranganathan, bibliothécaire indien de la première moitié du XXe siècle, est le premier à introduire la notion de facettes dans le domaine de la gestion documentaire. Garfield (1984) rapporte que l'idée de la théorie des facettes est venue à Ranganathan lors d'une *eurêka experience* réalisée en 1924 alors qu'il assistait à une démonstration de Légos dans un grand magasin londonien. Les Légos étaient mélangés mais il en existait différentes sortes : roues, portes, mur, fenêtres peuvant être assemblés pour construire différents objets. Cette métaphore du jeu de Légos est courante pour comprendre la classification. Ainsi l'exposition consacrée à Linné à Goteborg en été 2007 commence par une salle dans laquelle le visiteur dispose d'un bac contenant des Légos qu'il peut assembler à sa guise.

Les principes de la théorie des facettes sont contenus dans la classification Colon (Ranganathan, 1933) La théorie des facettes examine des problèmes complexes, les décompose en petites pièces (analyse) et assemble ces pièces ensemble. Cette méthode est connue sous le nom de « méthode analytico synthétique ». Ranganathan considère que tout document peut être décrit par une facette principale, représentant le domaine de la connaissance auquel on peut le rattacher (agriculture, médecine..) et 5 autres facettes correspondant aux

5 aspects de la réalité : Personnalité, Matière, Energie, Espace, Temps. Ranganathan voit dans ces cinq facettes une formule universelle permettant de classer chaque objet de connaissance. Cette formule est connue sous l'acronyme de PMEST où :

- P désigne la Personnalité : l'objet dont il est question (personne ou évènement dans une classification historique, un animal dans une classification en zoologie);
- M désigne la Matière, en quoi est fait l'objet d'étude
- E désigne l'énergie, représente l'opération, l'action décrite par rapport à l'objet
- S désigne l'espace dans lequel s'insère l'objet
- T désigne le temps

Ces 5 dimensions fondamentales et universelles de toute connaissance sont à rechercher dans les fondements de la philosophie Hindouiste (Couzinet, 2006).
Chaque facette est décrite par un code comme une langue est organisée autour d'une syntaxe. La formule PMEST propose une syntaxe dans cet ordre : Sujet ou objet (Personnalité), Verbe (Energie), Complément de manière (Matière), Complément de lieu (Espace), Complément de temps (Temps). Un document est donc décrit par une succession de 5 codes toujours dans le même ordre de telle sorte qu'un document ait une place et une seule dans la bibliothèque.
Là où les classifications hiérarchiques doivent prévoir tous les possibles, la classification par facettes combine entre elles les modalités des facettes autorisant la création de combinaisons correspondant à des éléments nouveaux qu'il aurait été impossible de classer dans des catégorisations a priori.

Nous allons nous pencher sur les travaux qui, à la suite de Ranganathan, ont permis d'expliciter, de clarifier le contenu de la théorie des facettes. Ces travaux ont été surtout le fait de l'école indienne de classification et du Classification Research Group (CRG).

b) Les successeurs de Ranaganathan

La théorie des facettes s'est développée suite aux travaux de Ranganathan dans trois écoles de pensées (La Barre, 2004) : indienne, américaine et anglaise :

- L'école indienne correspond aux travaux du Library Research Circle,
- L'école anglaise correspond aux travaux du Classification Research Group (CRG) dirigé par Wells et Vickery,
- L'école Nord américaine connue sous le nom de Classification Research Study Group est dirigée par Phyllis Richmond et Pauline Atherton Cochrane.

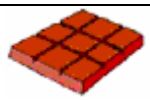
L'école qui a donné lieu aux travaux les plus conséquents est l'école anglaise du CRG. Ces travaux se sont distingués des travaux du père fondateur à deux niveaux : d'une part parce que l'analyse par facettes est désormais appliquée à des domaines pointus. D'autre part parce que les facettes sont considérées comme outil de classification et aussi outil de navigation.

A la suite des travaux de Ranganathan, des chercheurs se sont un peu éloignés de la formule PMEST qui malgré son caractère pédagogique et son universalité est difficile à mettre en œuvre. Dans les travaux ultérieurs, la théorie des facettes ne s'envisage plus de façon universelle mais contextuelle dans un domaine (Foskett, 1964). Comme le souligne Ellis (1999), cette tendance à appliquer la classification par facettes à des domaines pointus ne laisse pas augurer de perspectives très prometteuses lorsqu'il s'agira d'envisager la transposition de cette approche à un moteur de recherche web.

Broughton (2005) considère que les temps ont bien changé depuis les années 1930. A l'époque de Ranganathan, l'objectif d'une classification est l'organisation physique d'une collection papier, le rangement d'ouvrages dans une étagère. De ce point de vue, l'atout de la classification par facettes est sa capacité à ranger de façon linéaire des ouvrages classés de façon multidimensionnelle : de manière traditionnelle, lorsqu'on décrit un phénomène autour de plusieurs dimensions, on représente ce phénomène dans l'espace ou sur un plan si on est capable d'effectuer une projection en espace réduit. Dans le cas qui nous préoccupe, chaque ouvrage d'une bibliothèque est défini par des modalités appartenant à différentes facettes. Un codage rigoureux définit un ordre entre ces différentes modalités et permet de positionner un document à un seul endroit. Ceci est réalisé grâce au caractère « analytico synthétique » de la méthode qui consiste à déconstruire (analyse) et à reconstruire (synthèse) ensuite. Dans l'ère numérique, les contraintes de stockage n'existent plus.

L'avantage de la classification par facettes réside aussi dans le fait qu'elle facilite la navigation de l'internaute. En effet, d'après Kuronen et Pekkarinen (1999), les facettes permettent de mieux faire correspondre offre et demande d'information. Le marché de l'information se caractérise en effet par une offre d'information hétérogène : l'information officielle côtoie par exemple l'information informelle. La demande d'information correspond de son côté à des besoins différents. L'approche par facettes peut permettre de filtrer l'information dont l'utilisateur a besoin. La diversité des auteurs et des besoins renforce l'utilité de l'approche par facettes pour rendre ces deux logiques compatibles.

Taylor (1999) parle des facettes comme de l'ensemble des « *clearly defined, mutually exclusive, and collectively exhaustive aspects, properties, or characteristics of a class or specific subject* ». Les facettes sont orthogonales les unes les autres. L'orthogonalité signifie que chaque facette existe indépendamment des autres facettes. Nous pouvons reprendre l'exemple de Travis Wilson (2006) illustré *Figure 20*. Intéressons nous à des desserts. Nous avons choisi de les classer selon deux facettes : parfum et type de confection.

	 *Cerise*	 *Chocolat*	 *Noix*
 *Glace*	*5*	*2*	
 *Cookie*		*3*	*6*
 *Gâteau*	*4*	*7*	*1*

MENU :

1- gâteau aux noix
2- glace au chocolat
3- cookie au chocolat
4- gâteau à la cerise
5- glace à la cerise
6- cookie aux noix
7- Gâteau au chocolat

Figure 20: illustration de la notion de facette d'après Travis Wilson

Cette représentation didactique avec deux facettes permet de rendre compte de façon matricielle d'une réalité qui est beaucoup plus difficile à appréhender dès

que le nombre de facettes augmente. L'espace est alors multidimensionnel. Lorsqu'on recherche un dessert, on peut soit balayer les types de gâteaux puis les parfums soit les parfums puis les types de gâteaux.

On pourrait vouloir un gâteau au chocolat et aux noix mais en vertu du principe de mutuelle exclusion des éléments au sein d'une facette, cela est impossible. Si on veut un gâteau aux noix et au chocolat alors il ne faut pas choisir une facette parfum. On doit considérer une facette chocolat et une facette noix mais cela augmente le nombre de facettes avec pour chacune une modalité binaire (oui ou non ça contient du chocolat). Ceci revient alors au *tagging*.

Le tagging consiste à avoir un certain nombre de tags présentant chacun des modalités binaires. On aura ainsi un tag chocolat versus non chocolat. Ainsi la ressource « gâteau au chocolat et aux noix» est décrite par un trois tags : gateau ou non, chocolat ou non, noix ou non. On peut alors parfaitement concevoir un gâteau au chocolat et aux noix.

Pourquoi préférer un modèle à facettes plutôt qu'un modèle basé sur les tags ? Une des raisons est que le modèle à facettes comporte des règles de séparation et de combinaison des éléments à classer. Cela limite donc les combinaisons possibles et restreint le champ des possibles.

Cet exemple simple illustre toute la difficulté dans le choix des facettes. Certains auteurs se sont interrogés sur les principes de catégorisation des dimensions fondamentales d'un élément à décrire :

- L'école CRG (Vickery – 1963) suggère de recourir au dictionnaire qui, pour un terme donné, va permettre de remonter d'un niveau de généralisation (châtaigner – arbre). En effectuant des recherches successives à partir des concepts obtenus, on généralise de plus en plus jusqu'à trouver la propriété fondamentale de l'objet à décrire. Si on raisonne sur un ensemble de mots clés descripteurs d'un domaine, il est alors possible de regrouper ces éléments génériques en sous groupes cohérents qui constituent autant de facettes.
- Kyle (1957) précise que la classification dans les sciences sociales peut être facilitée par l'analyse de la citation. En effet, il peut être difficile de définir un terme mais plus facile d'établir des proximités entre des termes sur la base des citations identiques qu'ont ces termes.

D'autres travaux ont choisi des définitions des facettes beaucoup moins rigoureuses en confondant théorie des facettes et analyses pluridimensionnelles des données. D'autres parlent de facettes pour désigner la caractérisation d'un document par des propriétés particulières en terme de langue ou de genre par exemple. Or cela revient à dévoyer un peu la théorie des facettes. En effet, la théorie des facettes n'est pas simplement la recherche de points de vue sur une réalité à décrire mais la recherche d'un nombre limité de dimensions fondamentales qui caractérisent ce document. Telle est en tout cas la définition retenue par Maniez (1999) :

Facette : « gamme d'attributs communs fondamentaux en nombre limité utilisé comme technique d'analyse et de classification des concepts et des sujets ».

C'est la définition que nous allons retenir. Une facette n'est pas une simple variable qui peut prendre plusieurs modalités : c'est un attribut fondamental du document.

4. Web et classification

Il existe deux grandes façons de rechercher de l'information sur internet : les moteurs de recherche et les annuaires de recherche. Ce découpage, même s'il tend à s'estomper avec le temps, permet de rendre compte de deux influences scientifiques potentielles. La partie recherche d'information par mot clé dans les moteurs de recherche est liée au courant de « recherche d'information ». La partie « recherche d'information » dans les annuaires correspond à une utilisation de la classification dans le monde du web. Mai (2004) souligne qu'autant le développement des moteurs de recherche a pu bénéficier de transposition au monde du web d'approches théoriques développées dans le cadre du courant anglo-saxon *information retrieval* (on se souvient de PageRank de *Google* inspiré de l'analyse de la citation de la base SCI), autant les annuaires de recherche sur le web n'ont pas bénéficié de l'apport théorique des travaux de la communauté de la classification bibliographique. Il y a plusieurs raisons à cela, l'une d'elles étant que les travaux de recherche de la communauté de la classification ont porté sur des classifications appliquées au

domaine scientifique alors que l'information web n'est pas réduite à l'information scientifique. Mai souligne tout le potentiel et toute la richesse qu'il y aurait à enrichir les annuaires web avec une approche développée par le courant théorique de la classification. Un certain nombre de travaux se sont intéressés à la transposition au web de techniques de classification traditionnelles. Des travaux concernent aussi la transposition au web de systèmes de classification par facettes.

a) Web et classifications traditionnelles

Koch et al. (1997) montrent l'intérêt de l'approche classificatoire dans la recherche d'information. Le fait d'être capable d'adresser à une page web un élément d'une classification est intéressant à plusieurs titres :

- Cela permet d'une part une meilleure interopérabilité entre les ressources web. Si on est capable d'affecter à un document web plusieurs codes de classification correspondant par exemple à un code d'une classification brevet, à un code de la classification Dewey ou à un code Compass, alors on crée des passerelles, comme l'a montré Faucompré (1997), entre ces dimensions techniques économiques et scientifiques. L'idée alors est d'arriver à coder un même document avec divers systèmes de codage afin d'établir des passerelles entre le Web et les autres sources d'information.
- L'adoption d'une classification internationale permet un accès multilingue aux données.
- En dernier lieu, une classification internationalement reconnue permet de coder les données dans un format qui sera parlant pour les professionnels du domaine.

On peut donner quelques exemples d'organismes qui utilisent des classifications traditionnelles pour hiérarchiser le contenu de leur site web. La bibliothèque nationale du Canada[63] utilise la classification Dewey. Cyberstacks[64] utilise la classification de la bibliothèque du congrès pour catégoriser des ressources web dans le domaine de la science et de la

[63] http://www.collectionscanada.ca/caninfo/esub.htm
[64] http://www.public.iastate.edu/~CYBERSTACKS/

technologie. La classification de la NLM[65] est utilisée par Intute[66], catalogue de ressources internet validées dans le domaine de la médecine et de la santé.

Affecter un code classificatoire à un document demande de l'expertise humaine ce qui limite l'ampleur du corpus traité. Il n'y a pas d'exemple de corpus de bibliothèque classé par une méthode purement automatique. L'affectation manuelle peut être facilitée par une affectation pré automatique qui peut se faire grâce aux outils de la linguistique computationnelle (Godby, 2001). Ces méthodes comparent le langage naturel utilisé dans le document indexé (la page web) avec le système de classification. Dans ce domaine, le projet le plus abouti et encore en vigueur est Scorpion[67] Ce projet en *open source* dépend d'Online Computer Library Center qui affecte un code de la classification Dewey Decimal à une page web,

b) Web et classification par facettes

Si on transpose au monde du web les réflexions faites dans le cadre de la classification par facettes, on peut observer que l'expression du besoin par un internaute lors d'une recherche d'information sur le web s'exprime à travers un ou des mots clés définissant le thème de la recherche. L'identification du besoin de l'internaute se fait de façon très parcellaire et donc très appauvrissante. Le moteur de recherche sélectionne, à partir de ce qu'il a compris du besoin de l'internaute, un petit nombre de réponses qu'il présente de façon privilégiée, écartant une masse de données à laquelle l'internaute n'aura pratiquement jamais accès.

La théorie des facettes repose sur la prise en compte de plusieurs visions du monde. Elle considère qu'un document peut être décrit à travers de nouvelles dimensions qui ne se réfèrent pas seulement à son contenu thématique. Cette perspective ouvre des pistes très prometteuses aux moteurs de recherche. En effet, on abandonne l'indicateur de pertinence universel du moteur de recherche (à la *Google*) et on évolue vers une boîte à outils caractérisant un document web par différentes facettes, chacune pouvant être mobilisée par l'internaute lors de sa recherche d'information.

---

[65] National Library of Medecine

[66] http://www.intute.ac.uk/healthandlifesciences/medicine/

[67] http://www.oclc.org/research/software/scorpion/default.htm

Kim et al. (2006) caractérisent les ressources web à travers 6 facettes assez proches dans l'esprit de la définition de Ranganathan. Chaque facette inclut des sous facettes chacune ayant diverses modalités. Il propose le modèle suivant :

1. sujets (chose concrète, entité ; action et procédé ; type d'agent ; récence ; finalité et fonction ; universitaire / populaire
2. applications (fonction ; en ligne/en différé ; payant/gratuit ; transmission à sens unique/multiple ; produits/services)
3. localisations (découpage physique ou géographique ; découpage politique ; découpage économique ; degré de ruralité ; échelle)
4. médias/types/supports/outils (forme de publication, type de document de référence, forme de l'information, [formel/informel]; fonctions ; médias)
5. organisations (officiel/non officiel ; fonction ; avec ou sans espace physique)
6. personnes (âge, culture/ethnie ; niveau de formation ; niveau d'expérience ; tendance sexuelle ; profession ; statut ; groupes ; rôle)
7. information générale (champ d'application)
8. langues (découpage géographique)

Des chercheurs chinois (Xiaochuan et al., 2007) ont présenté un prototype disponible à l'adresse http://infoset.apexlab.org. Il permet de classer des pages de blogs sur un continuum page informative – page affective. Une capture d'écran de ce prototype est présentée *Figure 21*. Une page affective est une page de blog sur laquelle l'auteur exprime ses sentiments, ses pensées, ses émotions. Les pages informatives sont orientées vers des sujets. Les auteurs précisent qu'ils ont observé 85% de pages de blog affectives et 15% de pages de blog informatives. Les auteurs enrichissent donc la recherche par sujet d'une facette correspondant au caractère informatif de la page.

Figure 21 : Interface du prototype http://infoset.apexlab.org

Bast et Weber (2006) décrivent un prototype de moteur de recherche à facettes dans le domaine de la recherche d'information. Ces auteurs s'éloignent de la définition des facettes retenue par les auteurs antérieurs. Il ne s'agit plus de considérer les facettes comme décrivant les propriétés intrinsèques des documents répondant à une requête. Les auteurs se servent des facettes pour gérer l'expansion de requête : lorsque l'internaute tape une requête, le système affiche à la volée une facette appelée *refine by word* qui va reprendre les requêtes commençant par le texte de la requête. Ainsi on observe dans le prototype présenté Figure 22 qu'en commençant à taper « intellige », l'ordinateur génère une liste de requêtes commençant par « intellige ». Ces requêtes constituent la première facette. Dans un tel système, l'analyse par facettes permet de décrire des documents répondant non pas à une seule requête comme dans les cas précédents mais à un ensemble de requêtes qui correspondent à des termes qui ont le même radical que le terme de la requête.

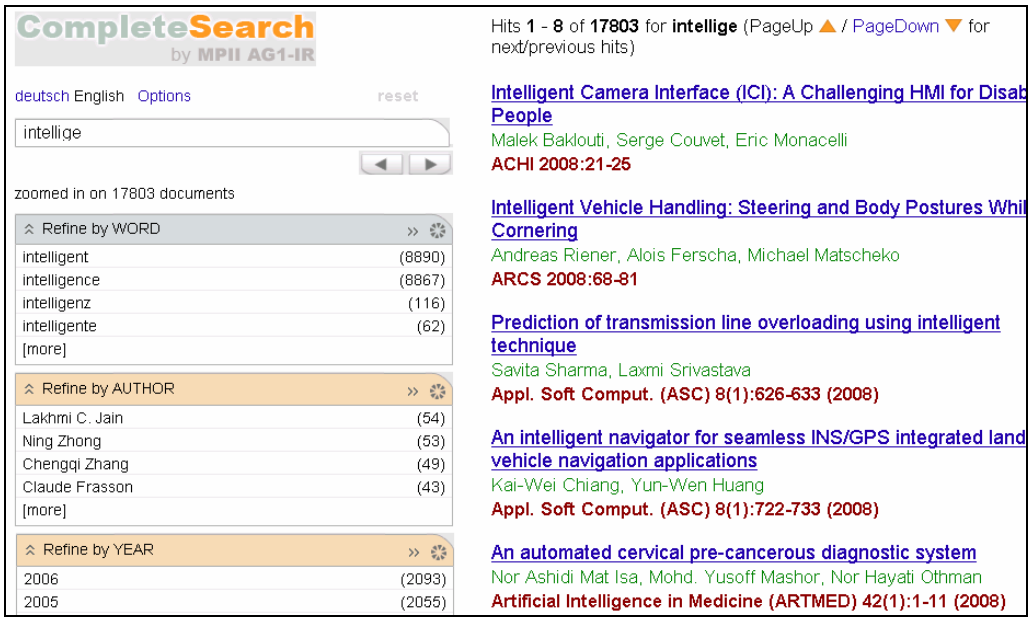


Figure 22 : Interface du prototype construit par Bast et Weber (2006)

5. Positionnement de notre approche par rapport à l'état de l'art

De toutes ces définitions des facettes, nous avons retenu celle de Maniez (1999) : « *gamme d'attributs communs fondamentaux en nombre limité utilisé comme technique d'analyse et de classification des concepts et des sujets* ». La recherche d'information sur internet, par son caractère démocratique tant du point de vue de la création des contenus que de leur utilisation soulève de nouveaux enjeux auxquels les facettes traditionnelles développées en « recherche d'information » ne permettent plus de répondre.

- Les grandes classifications sont utilisées pour classer des documents de type scientifiques, techniques ou académiques. La question de leur validité ne se pose pas. Sur internet, on doit faire face à une information qui n'est pas forcément validée : l'information web peut comporter des erreurs, des biais, des éléments subjectifs. La question de la validation de l'information web est un enjeu important.
- Dans la logique scientifique pour laquelle les classifications ont été déployées, la question de la personne qui accueille l'information ne se pose pas. Sur internet au contraire, le besoin peut être exprimé par une entreprise, un scientifique, un curieux, un enfant. Là encore une facette

> qui s'intéresserait au profil du chercheur d'information pourrait être intéressante. Le niveau de langage ne sera pas le même selon que le document s'adresse à un enfant, ou à un adulte, à un scientifique ou à une personne qui recherche un document de vulgarisation.

Ce n'est plus uniquement le contenu thématique qui nous intéresse dans un document mais la façon dont ce contenu est présenté : le niveau de subjectivité du discours, le niveau de langage du document. Nous privilégions donc l'identification de facettes de formes qui sont tout aussi fondamentales dans le contexte web que des facettes de contenu.

Cette approche correspond un peu à la notion de *common isolated* de Ranganathan (1933). Les « isolats communs » concernent la forme du document. Ils ne définissent pas le sujet à proprement parler : ce sont plus des auxiliaires. Le genre d'un document (dictionnaire, manuel, bibliographie, atlas, support visuel, biographie, périodique) est un exemple « d'isolat commun » aussi appelé « facette de forme ». (Shah et Kumar – 2006). La grosse différence qui existe entre la notion de « facette de forme » et notre acception des facettes est que chez Ranganathan et ses successeurs, ces « facettes de forme » sont des auxiliaires. Nous considérons qu'elles sont fondamentales si on se met du point de vue de l'internaute qui effectue une recherche d'information. Notre objectif est de renverser la focale et de passer d'une focale centrée sur le document à une focale centré sur l'utilisateur.

B. Les indicateurs de pertinence

1. La pertinence en recherche d'information

Il est intéressant d'analyser les perspectives de recherche dans le domaine de la pertinence des moteurs de recherche à la lueur des travaux scientifiques qui ont été réalisés dans le domaine de la pertinence en « recherche d'information ». Ces travaux peuvent en effet servir de cadre méthodologique pour comprendre les évolutions émergentes actuelles et envisager des perspectives renouvelées dans le domaine de la pertinence des moteurs de recherche.

La question de la pertinence est au centre de la « recherche d'information ». En effet, tout processus de recherche d'information est orienté vers la réponse à un besoin qu'il s'agit de satisfaire au mieux. D'où la nécessité d'introduire des mesures de la satisfaction de ce besoin.

Le processus de recherche d'information est décomposé en un certain nombre d'étapes. La compréhension de ces étapes est importante pour appréhender les multiples éclairages qui ont pu être apportés à la notion de pertinence. Le schéma de Simmonot (2002) présenté *Figure 23* est utile pour décomposer ce mécanisme. La première étape consiste en la perception d'un problème d'information. Ce problème va devoir être formalisé et traduit dans un langage compatible avec celui de la machine. Kislin (2007), dans sa thèse, aborde le cas particulier de l'intelligence économique dans lequel on retrouve deux acteurs : le décideur qui a un problème décisionnel et le veilleur qui a un problème informationnel. Le veilleur a au départ un rôle de médiateur pour traduire le problème décisionnel en problème informationnel. L'auteur propose alors une définition de l'intelligence économique comme « démarche collaborative de compréhension et de résolution de problèmes décisionnels et informationnels ». La requête est alors récupérée par le moteur qui va consulter sa collection et extraire la liste des documents pertinents qui seront hiérarchisés. Ceux-ci sont alors restitués à l'utilisateur qui va confronter son besoin d'information aux documents qui lui sont fournis pour créer de l'information et de la connaissance.

A chaque étape de ce processus, il existe des biais potentiels conscients ou inconscients : il existe par exemple une distorsion possible au niveau de la requête qui peut représenter plus ou moins bien le besoin d'information de l'internaute et à plus forte raison entre les documents restitués par le moteur et les besoins d'information de l'utilisateur.

Il est important de noter que sur ce schéma (*Figure 23),* chaque acteur possède une information imparfaite et qu'il a une visibilité réduite de l'ensemble du système. L'utilisateur maîtrise le processus jusqu'à la formalisation de la requête. L'outil de recherche maîtrise le processus à partir de la formalisation de la requête mais l'outil de recherche n'a pas de visibilité sur la partie préparatoire à la création de la requête et l'utilisateur peu de visibilité sur l'algorithme utilisé pour lui renvoyer des documents pertinents. Il est donc

difficile de qualifier et d'améliorer la pertinence de tout un processus lorsqu'on n'en maîtrise pas toutes les étapes. La *Figure 24* reprend ces étapes en constituant deux boites noires pour bien illustrer le peu de visibilité de chaque partie sur l'autre partie et la difficulté d'avoir une vision d'ensemble du problème.

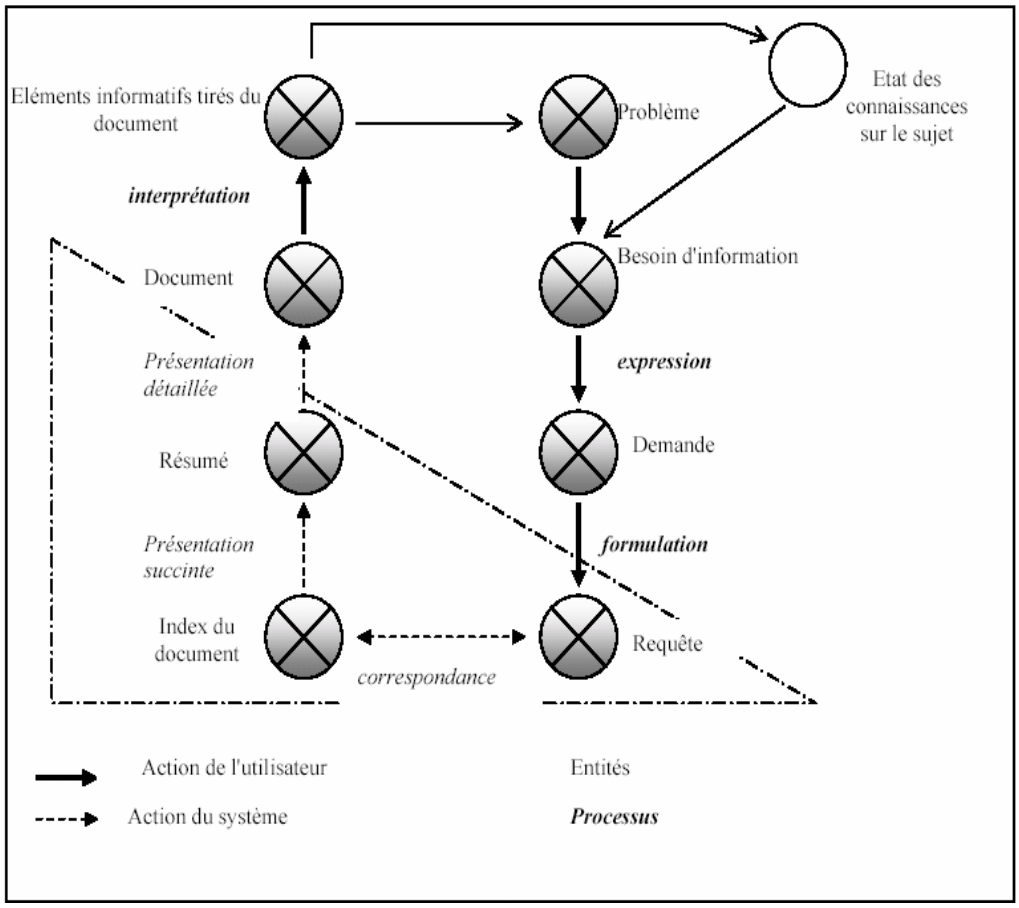


Figure 23 : Les principales étapes d'un processus de recherche d'information d'après Simmonot (2002)

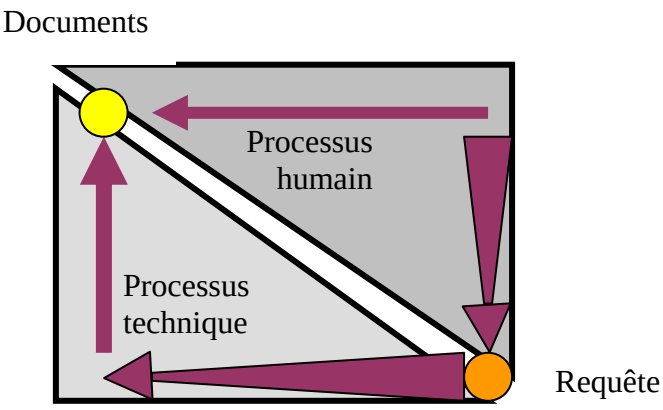


Figure 24 : Les principales étapes d'un processus de recherche d'information : deux boites noires

Il découle assez logiquement de cette vision du processus de recherche d'information que deux visions de la pertinence sont envisageables :

- La pertinence technique intègre la boite noire maîtrisée par le moteur de recherche. On parlera de pertinence à ce titre pour qualifier la relation de correspondance entre l'output (ensemble de documents) et l'input (requête formulée par l'internaute). Un document est pertinent par rapport à une requête pour Lavrenko (2004) s'il y a une correspondance sémantique substantielle entre le document et la requête. Cette pertinence est appelée *relevance* dans la littérature anglo-saxonne. On parle aussi de « pertinence algorithmique ». Cela correspond donc à une vision limitée de la pertinence restreinte à la partie que maîtrise l'outil de recherche. Lorsqu'il s'agit d'envisager non plus la pertinence d'un document mais l'efficacité du système de « recherche d'information », on introduit la notion agrégée de rappel et de précision (Van Rijsbergen – 1979). Le rappel correspond au rapport existant entre le nombre de documents pertinents trouvés et le nombre de documents pertinents dans la collection. Le silence est le complémentaire à 1 de cette valeur. Il correspond au pourcentage de documents pertinents qui n'ont pas été renvoyés à l'utilisateur. La précision correspond au rapport existant entre le nombre de documents pertinents trouvés et le nombre de documents renvoyés. Le complémentaire à un de la précision est le bruit c'est-à-dire le pourcentage de documents présents dans la réponse du moteur qui ne satisfont pas la requête. Dans cette vision technique de la pertinence, l'utilisateur n'est pas pris en considération. On s'aperçoit que ces deux concepts de rappel et de précision, très présents dans les réflexions en « recherche d'information » sont aussi très ancrés dans une vision technique de la pertinence. Certains auteurs ont proposé de parler de *document retrieval* plutôt que *d'information retrieval* .
- La seconde forme de pertinence s'entend au sens de l'utilisateur. On parlera de pertinence d'un document au sens de l'utilisateur pour qualifier la correspondance qu'il existe entre ce document et le besoin de l'utilisateur. Cette pertinence utilisateur est donc beaucoup plus englobante car elle couvre tout le processus décrit en amont. Toutefois cette

pertinence est également beaucoup plus subjective, contextuelle et délicate à mesurer. En effet, la caractérisation de la pertinence d'un document par l'utilisateur met en jeu divers facteurs comme l'expérience passée de celui qui effectue la recherche, l'objectif qu'il poursuit, l'état de ses préférences. Heine (2000 b) reprend les différentes appellations de cette forme de pertinence. On parle de pertinence situationnelle (Wilson – 1973) psychologique (Heine - 2000) ou subjective (Harter – 1992).

Mizzaro (1998) a proposé de modéliser la notion de pertinence autour de 4 dimensions. Dans ce modèle, les formes technicistes et humaines de la pertinence apparaissent aux extrémités d'un continuum complexe qu'il est intéressant de découvrir :

- La première dimension de la pertinence correspond au niveau de représentation du problème chez l'utilisateur. Au départ, le besoin d'information est flou et peu formalisé. Belkin et al. (1982) ont introduit la notion de *Anomalous State of Knowledge* (ASK) pour signifier que l'utilisateur n'est pas pleinement conscient de ce qu'il recherche. Ingwersen (1992) parle *d'Incomplet* ou *d'incertain state of knowledges*. Ce besoin sera alors exprimé, verbalisé de façon à être intelligible par un outil de recherche. Mizarro distingue alors le RIN (*real information need*), le PIN (*perceived information need*), la *request* et la *query*. Taylor (1968), dans une approche voisine, distingue le besoin viscéral, le besoin conscient, le besoin formalisé et le besoin compromis (pour rendre compte de l'arbitraire qu'il y a à enfermer un besoin de recherche exprimé dans un langage compatible et reconnu par un outil de recherche). La requête est une représentation de surface d'un concept abstrait.
- La seconde dimension principale de la pertinence est celle du niveau d'information. Le moteur de recherche restitue au départ une information composée d'une liste comprenant une vision synthétique des documents qu'elle contient (Mizzaro parle de *surrogate*). Le document apparaît à un deuxième niveau. Le troisième niveau correspond à l'information c'est-à-dire à ce que l'utilisateur du système va créer à partir de la lecture du ou des documents. On définit donc l'information comme un processus de création découlant de la rencontre du besoin et des documents.

- La troisième dimension est le temps. En effet, le processus de recherche d'information n'est pas linéaire mais en boucle. Les premières informations collectées conduisent à un affinage de la perception du besoin et à une redéfinition de la requête. Les documents déjà lus ont une influence sur l'intérêt que l'utilisateur va porter à des documents nouveaux. Un document redondant d'un document déjà consulté aura donc de ce fait une pertinence nulle.
- La dernière dimension est multiforme. Mizzaro parle de composant pour caractériser 3 dimensions : le sujet, l'objectif et le contexte (ce que je sais déjà, le temps que je suis prêt à y mettre, ce que je peux comprendre, le prix que je suis prêt à y mettre).

Si on reprend cette formalisation de Mizzaro, on peut décrire la pertinence technique des moteurs de recherche à partir des 4 dimensions :

- Niveau de représentation du problème chez l'utilisateur : requête
- Niveau d'information : page de résultat du moteur de recherche
- Temps : non pris en compte
- Composant (sujet / objectif / contexte) : seul le sujet est considéré

Si on considère toujours cette formalisation, la caractérisation de la pertinence humaine correspond à :

- Niveau de représentation du problème chez l'utilisateur : besoin
- Niveau d'information : information
- Temps : pris en considération
- Composant (sujet / objectif / contexte) : tous les trois sont considérés

Mizzarro a proposé une représentation graphique (***Figure 25***) pour rendre compte dans le plan des deux premières dimensions.

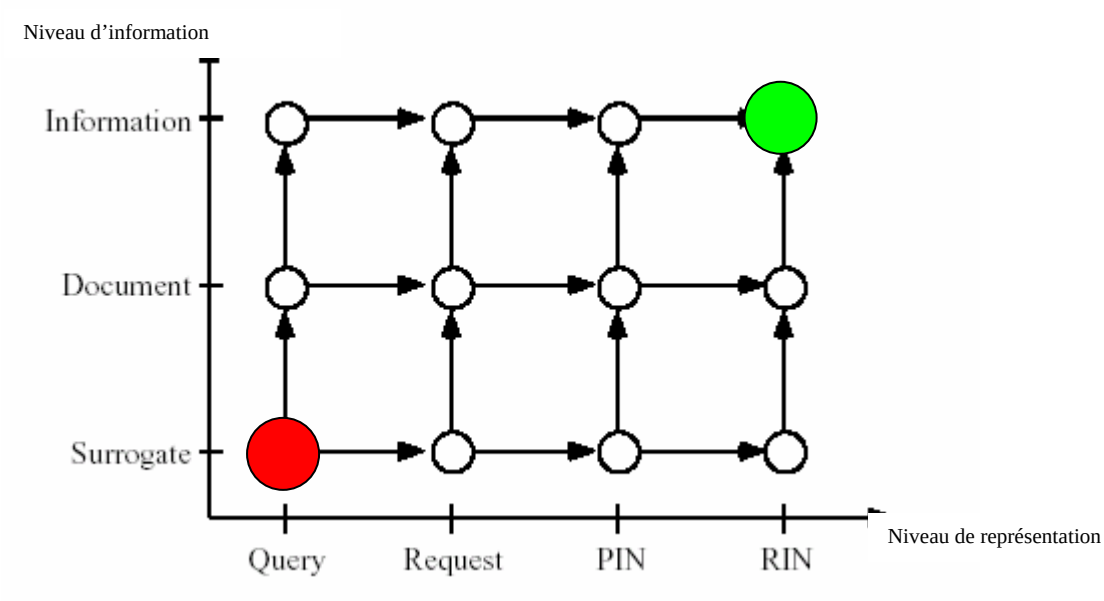


Figure 25 : la pertinence selon Mizzaro

Chaque définition de la pertinence y apparaît comme un point sur le plan. La pertinence technique correspond au point en bas à gauche du graphique. La pertinence avec vision utilisateur correspond au point situé en haut à droite. Il faudrait pouvoir ajouter deux axes supplémentaires pour prendre en compte la dimension temporelle et la dimension composant.

Ce graphe a le mérite d'être visuel et de proposer une démarche permettant d'évoluer d'une vision techniciste à une vision utilisateur de la pertinence. L'auteur propose un déplacement sur l'une ou l'autre des 4 dimensions. Ce glissement a fait l'objet de transposition dans le monde des moteurs de recherche, certaines de ces perspectives de glissement étant déjà mises en œuvre. En voici quelques exemples :

- Le traitement de la requête en langage naturel permet d'évoluer vers la droite de l'axe des abscisses.
- Le bouton « j'ai de la chance » de *Google* permet à l'internaute d'ouvrir le premier document par défaut correspond à une évolution sur l'axe vertical.
- Le profilage de l'internaute et la prise en compte du contexte de la recherche correspond à un déplacement sur l'axe 4

Ce balayage de la littérature dans le domaine des indicateurs de pertinence en « recherche d'information » nous conduit à observer une double évolution. La première traduit l'émergence d'indicateurs de pertinence centrés utilisateur. La

seconde évolution concerne la prise en compte du contexte de la recherche. Nous allons consacrer les deux paragraphes suivants à ces deux points.

2. Vers des indicateurs de pertinence de moteurs de recherche centrés utilisateur

Aujourd'hui une littérature scientifique importante est consacrée aux indicateurs de pertinence centrés sur l'utilisateur. L'objectif général de ces travaux est d'arriver à disposer d'informations destinées à enrichir les données fournies par la requête. Nichols (1997) précise deux voies pour y parvenir : solliciter l'internaute (même s'ils sont souvent assez réfractaires à toute forme d'explicitation) ou identifier ses préférences de façon implicite. La seconde voie consiste généralement à mobiliser les techniques du *web usage mining*. Il s'agit selon le cas d'exploiter les informations relatives à l'historique de la navigation de l'internaute ou de s'appuyer sur un groupe d'internautes qui auraient des préférences similaires.

Claypool et al (2001) se sont intéressés à l'identification des facteurs qui, dans la navigation de l'internaute, constituent des marqueurs de son intérêt pour une page web. Ces auteurs ont développé un navigateur (*curious browser*) qui capture, pour une page web donnée, les actions de la souris et du clavier. Grâce aux informations ainsi collectées, il est possible de construire différents types d'indicateurs :

- indicateurs de type marquage : combien de pages ont été mises dans les favoris, sauvegardées ou imprimées ?
- indicateur de manipulation : les ascenseurs ont-ils été utilisés, du texte a-t-il été copié - collé ?
- indicateur de navigation : des liens ont-ils été suivis ?
- indicateurs de répétition : la page a-t-elle fait l'objet de visites répétées ?

Tous ces comportements sur les pages sont comparés à une note mise par l'internaute qui a visité la page. L'objectif est d'étudier les facteurs les plus discriminants pour comprendre l'intérêt d'un internaute sur une page web. Il ressort de l'étude que le temps passé sur une page est un bon indicateur de l'intérêt d'une page. En connaissant mieux les pages qui ont plu à l'internaute, il est possible de reconstituer ses centres d'intérêts qui ne sont pas infinis. La

connaissance de ses préférences permet d'éclairer la requête actuelle et de l'enrichir.

La technique du filtrage collaboratif (*collaborative filtering*) peut également être mobilisée. Harlecker et al. (2004) présentent un état de l'art de cette technique. Le filtrage collaboratif considère que si un groupe d'internautes partage des intérêts similaires, alors les préférences d'un utilisateur peuvent être déduites au vu des préférences observées sur d'autres internautes qui lui sont proches. Ces techniques sont utilisées principalement dans les boutiques en ligne.

Nous allons chercher à mieux comprendre les motivations d'un internaute lors d'une recherche web. Pour cela, nous proposons d'organiser la présentation de cette section autour de trois étapes :

- Dans un premier temps, nous présenterons différentes typologies utilisées pour rendre compte de la motivation latente de l'internaute.
- Une seconde étape consistera à se demander s'il est possible de connaître automatiquement la motivation de la recherche.
- La troisième étape visera alors à proposer des pistes d'exploitation de l'information relative à l'objectif de la recherche pour évoluer vers un indicateur de pertinence « goal sensitive ».

a) Essai de typologie des motivations de recherche des internautes

Un certain nombre de recherches ont porté sur les motivations de l'internaute qui effectue une recherche sur le web. Broder (2002) distingue trois motivations principales qui se traduisent par trois familles de requêtes : « navigationnelles », « informationnelles » et « transactionelles »[68].

- Dans le cas d'une requête navigationnelle, l'intention de l'internaute est d'utiliser un moteur de recherche pour être conduit à une page web qui fait autorité sur une question. Il est parfois possible qu'il connaisse cette page web pour l'avoir déjà visitée mais que, n'ayant pas mémorisé son adresse

[68] Ce cloisonnement est articiel, un internaute pouvant avoir plusieurs motivations lorsqu'il effectue sa recherche d'information.

url, il recoure au moteur de recherche pour la retrouver. Dans ce cas, le processus de recherche est rapide et s'arrête lorsque l'internaute arrive sur la page qu'il avait en tête. Le bouton de *Google* « j'ai de la chance » correspond pleinement à la volonté de ce moteur de satisfaire les requête de type navigationnelle. *Google* propose, lorsqu'on clique sur ce bouton, de conduire directement l'internaute vers la première page web de sa liste. Au lieu de visualiser la page de données intermédiaires comportant la liste des pages web classées en mode abrégé, l'internaute accède directement à la page pertinente.

- Dans le cas d'une recherche informationnelle, l'intention de l'internaute est d'obtenir de l'information sur un sujet en consultant une ou plusieurs pages web. Le problème est alors peu défini ; il peut s'agir de générer des pistes pour des recherches futures.

La grosse différence qui existe entre requêtes navigationnelles et informationnelles, comme l'ont montré Uichin et al. (2005), est que dans le premier cas, l'internaute attend une réponse alors qu'il en attend plusieurs dans le second cas.

- Dans le cas d'une recherche transactionnelle, l'intention de l'internaute est d'atteindre un site web où une interaction est attendue (commerce, téléchargement, consultation de base de données). Rose et Levinson (2004) parlent de requêtes ressources pour désigner des requêtes débouchant sur la récupération de documents autres que textuels utilisables en ligne. Rose et Levinson distinguent plusieurs rubriques au sein de cette catégorie :
    - *Obtain goal*: l'objectif de la requête est de récupérer un document qui sera utilisé hors ligne : les paroles d'une chanson, une recette de cuisine
    - *Download*: l'objectif est de récupérer une ressource qui aura besoin d'être installée sur ma machine pour fonctionner
    - *Entertain* : un programme de jeu
    - *Interact* : l'objectif est d'obtenir sur le web un service dynamique

Après avoir défini les objectifs de la recherche sur internet, ces auteurs se sont intéressés à la mesure de la part respective de chacune de ces trois catégories. Ils ont pour cela mobilisé trois sources d'informations :

- La requête, en soi très éclairante, à laquelle il est généralement possible d'associer l'une ou l'autre des trois catégories sus citées. La requête « pubmed » par exemple est caractéristique d'une requête navigationnelle, l'internaute cherchant sans doute à se connecter à la base de données de Medline. Une recherche de type « algorithmes génétiques » est plutôt informationnelle car il existe plusieurs ressources dispersées sur ce sujet et la façon générique de présenter la question laisse penser que l'internaute souhaite plutôt balayer cette diversité.
- Les traces laissées par les internautes. Les moteurs de recherche peuvent reconstituer le parcours de l'internaute lors d'une recherche en exploitant les traces de leurs fichiers log. Ce parcours comporte des marqueurs qui correspondent de façon privilégiée à certaines catégories de requêtes. Une recherche courte correspond plus volontiers à une recherche de type navigationnelle. Une requête associée à des téléchargements correspond plus volontiers à une requête de type ressources.
- La troisième façon consiste à reconstituer les objectifs latents de la requête à partir d'une information explicitement demandée à l'internaute.

En utilisant l'une ou l'autre de ces sources d'information, Broder (2002) et Rose et Levinson (2004) sont chacun arrivés à quantifier en pourcentage la part respective de chacune de ces requêtes. Les résultats de Rose et Levinson sont présentés dans le *Tableau 16.*

Type de requête	Pourcentage estimé par Rose et al.
requêtes navigationnelles	15%
requêtes informationnelles	60%
requêtes ressources	25%

Tableau 16 : répartition des requêtes d'après Rose et Levinson

Broder qui a privilégié une information obtenue à la suite d'un questionnaire trouve des résultats sensiblement différents, présentés *Tableau 17*.

Type de requête	Pourcentage estimé par Broder.
requêtes navigationnelles	25%
requêtes informationnelles	40%
requêtes transactionnelles	35%

Tableau 17: répartition des requêtes d'après Broder

Jusqu'à présent, nous avons montré l'intérêt qu'il y avait à transposer au monde des moteurs de recherche des travaux menés dans le cadre de la « recherche d'information ». Cette transposition n'est envisageable que dans le cas où l'internaute recherche de l'information. Or on se rend compte que cette problématique de recherche d'information se retrouve dans au mieux 60% des cas. Cela signifie qu'il faudra mettre en œuvre d'autres approches pour comprendre les indicateurs de pertinence des moteurs de recherche.

b) Déterminer automatiquement l'objectif d'une recherche

Certains travaux récents se sont intéressés à l'identification automatique de l'objectif de l'internaute (Uichin et al., 2005). Ce problème est délicat : il pose la question de la prédictibilité de l'objectif de la requête sans poser la question à l'internaute. Pour simplifier leur analyse, ces auteurs ont raisonné uniquement sur les requêtes informationnelles et navigationnelles.

Uichin et al. considèrent au départ un ensemble de 50 requêtes faisant partie du top 50 de *Google*. Pour chacune d'elles, ils demandent à des internautes de préciser s'ils considèrent plutôt cette requête comme informationnelle ou navigationnelle. Si 80% au moins des internautes interrogés sont d'accord sur le statut d'une requête, alors on considère qu'il y a consensus. Sur l'échantillon considéré, les auteurs arrivent à un consensus sur ce type de requête dans 60% des cas. Les 40% autres sont souvent des cas de requêtes relatives à l'informatique ou à des personnes. Ces dernières requêtes peuvent être reconnues automatiquement par une machine sur la base d'un certain nombre de mots clés. Les auteurs s'aperçoivent que dans les 40 % des requêtes

non prédictibles, on observe que l'expérience passée de l'internaute interrogé a une influence sur la détermination de l'objectif de la requête. Dans le cas d'une requête sur une personne, un enquêté qui connaît cette personne aura tendance à rechercher une page spécifique (requête navigationnelle) là où le débutant cherchera plutôt à s'entourer de plusieurs réponses (requête informationnelle). On s'aperçoit donc par ce détour que l'objectif d'une recherche fait intervenir des critères personnels et qu'en fonction de son expérience passée, l'internaute poursuivra un ou l'autre des objectifs génériques définis précédemment.

Les auteurs se sont ensuite attachés aux 60% de requêtes prédictibles humainement pour voir dans quelle mesure ces requêtes pouvaient voir leur statut déterminé automatiquement. Pour cela ils ont utilisé deux sources d'information :

- l'analyse des clics passés par les internautes. Les auteurs ont analysé le comportement d'internautes s'étant intéressés à des requêtes identiques par le passé. Cette information peut être connue par les moteurs de recherche qui conservent, dans leur fichier « .Log » la trace des requêtes et des parcours des internautes. Si une requête donnée A donne lieu à la visualisation par les internautes d'un petit nombre de réponses, alors il s'agit sans doute d'une requête navigationnelle. Si au contraire, la requête donne lieu à un grand nombre de pages visualisées, alors la requête est plutôt informationnelle. Le nombre de pages différentes visualisées par des internautes pour une requête donnée est donc un indicateur de son caractère plus ou moins informationnel. Les données sont statistiquement consolidées sur un grand nombre d'internautes ayant réalisé cette requête par le passé. Le fait que les requêtes privilégiées fassent partie du top 50 des requêtes de *Google* permet d'avoir un grand nombre d'utilisateurs et des résultats statistiquement pertinents. Pour une requête, on peut définir un indicateur correspondant par exemple au nombre moyen de pages différentes visualisées par les internautes. Cela permet d'en inférer l'objectif de la requête pour l'internaute moyen. La limite principale de cette approche est qu'elle s'intéresse à l'utilisateur moyen. Elle ne se situe pas au niveau de l'internaute qui pose la question mais de tous les internautes qui ont posé la question par le passé. Par une telle analyse, on

s'éloigne d'un niveau de personnalisation plus fin qui aurait fait intervenir l'expérience passée de l'internaute, lui-même, sur ce type de requête. Au lieu d'analyser les traces des internautes ayant posé ce type de requête, il aurait été possible d'analyser le parcours de l'internaute en analysant son comportement dans son historique de navigation. Cette analyse de l'historique de l'internaute permettrait d'avoir un niveau d'analyse plus fin et de mieux cerner l'objectif de la recherche. Cette trace de l'historique est délicate à mettre en œuvre. Elle correspond à la notion de profil utilisateur qui comporterait dans notre cas pour chaque requête passée de l'internaute son type estimé à partir du nombre de pages qu'il a visualisées lors de cette requête. Cette mise en œuvre pose des problèmes de respect de la personne par la traçabilité qu'elle présuppose. Elle est donc difficilement implémentable en l'état.

- Pour connaître le type de requête (navigationnelle, informationnelle), les auteurs se sont également intéressés au texte figurant sur les liens hypertextuels[69]. Pour analyser le type de chaque requête, ils récupèrent un corpus de liens hypertextes qui comportent le texte de la requête sur le lien hypertexte. Ils observent ensuite, pour chaque lien, vers quelle page il pointe. Si ces liens pointent vers un nombre limité de pages web, cela signifie qu'il existe sur le sujet une page de référence et la requête est qualifiée de requête navigationnelle. Si les liens pointent vers un grand nombre de pages différentes, alors la requête est informationnelle. Dans cette logique, ce n'est donc pas l'utilisateur qui détermine le statut d'une requête mais la variété des réponses proposées par le moteur de recherche. Ca ne correspond donc pas forcément à l'objectif latent de l'internaute. En raisonnant ainsi, on fait l'hypothèse que si la requête de l'internaute renvoie à une page qui fait autorité, cela signifie que l'internaute recherchait une page autorité. Ce type de raisonnement donne du poids à des algorithmes de type Clever (hub et authority) (Chakrabarti et al (1999) qui sont capables de définir la notion d'autorité d'une page web.

[69] texte souligné en bleu lorsqu'on est sur une page web, point de départ du lien hypertexte

Grâce à ces deux sources d'information, les auteurs parviennent à retrouver de façon automatique le statut des requêtes dans 90% des cas.

c) Vers un indicateur de pertinence *Goal sensitive*

Il est important pour le moteur de recherche d'identifier le plus en amont possible à quelle catégorie se rattache une requête. Cela lui permettra de réorganiser ses réponses pour satisfaire le besoin de l'internaute. Dans le cas d'une requête navigationnelle, il s'agit de privilégier une page qui fasse autorité. Dans le cas d'une requête navigationnelle, il s'agit de faire ressortir plusieurs pages représentant la diversité des réponses possibles. Dou et al. (2007) confirment les résultats obtenus par Chirita et al. (2005) qui montrent que la personnalisation est efficace dans le cas de requêtes ambiguës. Tan et al. (2006) distinguent les requêtes fraîches des requêtes récurrentes. Les auteurs montrent que l'historique récent est utile pour les requêtes dites fraîches alors qu'un historique de navigation plus ancien est nécessaire pour les requêtes récurrentes. Teevan et al. (2005) montrent que la personnalisation est pertinente dans le cas de requêtes pour lesquelles on observe de fortes variations entre internautes sur les pages qu'ils trouvent les plus pertinentes. Ces pages sont dites à *large click entropy*. Par contre, ces techniques de personnalisation sont peu efficaces dans le cas de recherche navigationnelles où les usagers sont unanimes pour privilégier un petit nombre de pages parmi celles qui sont renvoyées par la requête.

3. Comprendre le contexte de la recherche d'information

Beaucoup de méthodes en « recherche d'information » reposent sur l'hypothèse d'indépendance de la pertinence entre les documents. Selon Heine (2000b), la pertinence d'un document est une donnée inhérente au document. Il est alors possible de calculer la pertinence de chaque document candidat et de les hiérarchiser ensuite par pertinence décroissante. Cette vision est retenue par les moteurs de recherche. Leur classement correspond au classement des pages par ordre décroissant de pertinence. Les méthodes utilisées dans les

comparatifs des moteurs de recherche utilisent toutes aussi une évaluation des pages web indépendamment les unes des autres.

Le mécanisme de la pertinence est plus complexe lorsqu'on fait intervenir le contexte de la recherche. En effet, le document visualisé par l'internaute, à un instant t, aura une pertinence qui dépendra des documents antérieurs que cet internaute a pu visualiser dans un passé plus ou moins éloigné. La pertinence d'un document se définit donc de manière contextuelle c'est-à-dire par rapport aux autres documents contenus dans le corpus de données. Plusieurs niveaux contextuels peuvent alors être identifiés :

- Le premier niveau est celui de la requête. Le moteur de recherche renvoie plusieurs résultats. La pertinence d'une page dépendra de l'ordre dans lequel l'internaute va parcourir les réponses qui lui sont proposées. Si l'internaute découvre par exemple la page web B après avoir découvert une page A très similaire alors la page B n'apportera que peu d'information nouvelle et aura donc une pertinence faible à ses yeux. Il en aurait été tout autrement si l'utilisateur avait commencé par visualiser la page B.
- Le deuxième niveau est celui de la recherche d'information. Une recherche d'information peut donner lieu, lorsque le problème est mal défini à la production de plusieurs requêtes au moteur de recherche. Les requêtes en effet s'affinent pour être plus proches du besoin d'information. La pertinence d'une page dépendra des pages visitées dans le cadre des autres requêtes issues de la même recherche. La encore une page redondante par rapport à une page visualisée antérieurement sera sans doute qualifiée de non pertinente.
- Le troisième niveau consiste à retrouver une information déjà identifiée par l'internaute dans le passé. Dans une requête de type navigationnelle par exemple, l'internaute pourra chercher à retrouver le nom d'une page web qu'il a découvert dans le passé. La redondance aura alors dans ce type de cas une vertu positive.
- Enfin le dernier niveau intègre toutes les connaissances que l'internaute a pu acquérir sur le web ou ailleurs. Cela contribue à son niveau de connaissance sur le sujet et influe sur la perception qu'il aura d'un document donné.

Le schéma de la *Figure 26* illustre ces multiples niveaux d'influence dans la perception d'un document. On doit considérer que la pertinence d'un document dépend non seulement de l'utilisateur mais aussi des documents antérieurement visualisés par cet internaute.

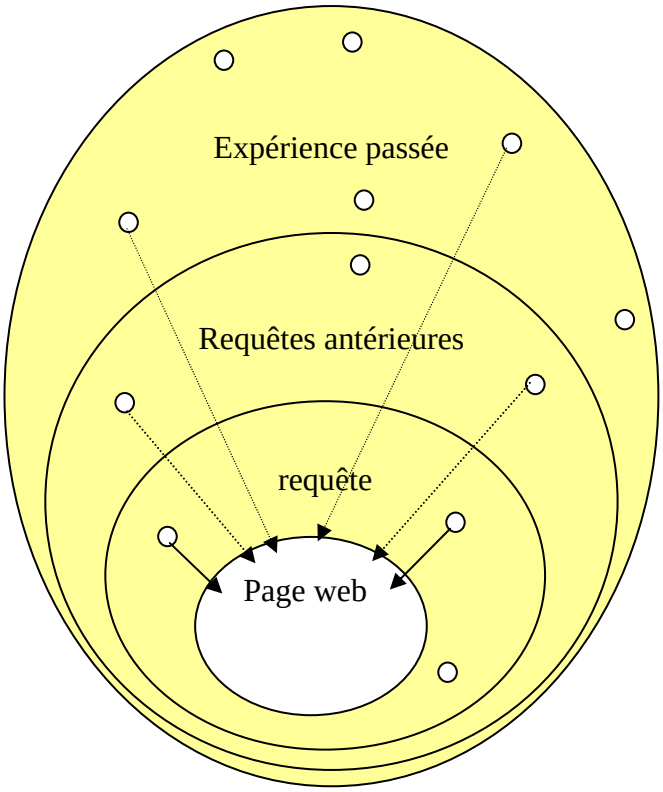


Figure 26 : La requête dans son contexte

La question est alors de savoir de quelle façon la pertinence d'un document peut être affectée par des documents antérieurs. Une revue de la littérature nous permet de voir que la notion de redondance est au cœur de cette interaction. Si un document est redondant par rapport à un document antérieur, alors, il aura une pertinence nulle. La redondance est donc négative dans cette acception. Goldstein et al. (1998) se sont attachés à modéliser la notion de pertinence en proposant de combiner pertinence et anti-redondance. Ils classent les résultats du moteur d'une part selon leur pertinence par rapport à la requête et d'autre part selon leur degré de nouveauté par rapport aux autres documents de la collection. Les auteurs proposent une mesure de la nouveauté d'un document à travers le calcul de la similarité textuelle entre le document courant et les documents antérieurement lus par l'internaute. Ils définissent le concept de pertinence marginale maximale (*Marginal Maximal Relevance*) qui est une

combinaison linéaire de la pertinence par rapport à la requête et la nouveauté. Un document a une pertinence marginale maximale s'il répond à la requête et contient peu de similarité avec les documents précédents.

Nous considérons pour notre part que la redondance peut être positive dans deux familles de requêtes :

- Dans le cas des requêtes navigationnelles, l'internaute cherche à retrouver l'adresse d'une page web qu'il a bien souvent déjà visualisée. Dans ce cas, qui correspond d'après Broder (2004) à une requête sur 4, la page qui est recherchée sera d'autant plus pertinente qu'elle correspondra à la page antérieurement recherchée.
- Dans le cas de requêtes informationnelles, la redondance signifie la confirmation. Lorsque l'internaute découvre un sujet qu'il ne connaît pas, il visualise plusieurs réponses parfois redondantes. Cette redondance est positive car elle va servir de validation subjective du crédit qui va être accordé à l'information. Plus un grand nombre de documents supposés indépendants va dans le sens d'une information et plus celle-ci aura de valeur aux yeux de l'internaute

Il existe d'autres éléments permettant de lier la pertinence d'un document aux documents découverts antérieurement. Pour mieux comprendre les phénomènes en présence, il est utile de faire un détour en présentant la notion de biais cognitifs et son application en recherche d'information

4. Biais cognitifs en recherche d'information

La recherche d'information sur le web peut être apparentée à un processus de stimulation fort qui doit être canalisé par l'internaute. L'internaute doit en effet capturer des stimuli externes, les analyser les uns les autres, les combiner avec des informations antérieures pour les transformer en connaissance et en savoir. De ce point de vue, la recherche d'information peut être analysée à la lueur de la théorie de la rationalité limitée du décideur (Simon, 1982). La surabondance de l'information se double d'une incertitude : elle conduit l'internaute à développer des mécanismes de protection inconscients appelés biais cognitifs. Ces biais ont pour objectif de simplifier la complexité de la réalité et éviter la

surcharge informationnelle. Ce réflexe est une nécessité vitale pour garantir la santé mentale de l'internaute. Plusieurs phénomènes sont à l'œuvre :

a) la dissonance cognitive (Festinger – 1957):

Un décideur, confronté à un problème, va concevoir une première estimation de solution. Ensuite, il affinera cette solution. Le principe de dissonance cognitive introduit par Festinger (1957) considère que « L'existence simultanée d'éléments de connaissance qui, d'une manière ou d'un autre, ne s'accordent pas (dissonance) entraîne de la part de l'individu un effort pour les faire, d'une façon ou d'une autre, mieux s'accorder (réduction de la dissonance) ».

Dans le contexte présent, on s'intéresse à la dissonance cognitive associée à la prise en compte par l'individu d'une information qui ne renforce pas la connaissance qu'il a accumulée sur le sujet. Dans un contexte d'information surabondante, l'internaute ne peut pas accorder autant de poids aux milliers de réponses qui lui sont renvoyées par le moteur de recherche. Celui-ci va donc se concentrer sur les premières pages et, au sein des premières pages, il va accorder beaucoup de poids aux toutes premières réponses. L'internaute a donc une limitation intrinsèque dans sa capacité à absorber la nouveauté. Il va accorder beaucoup de poids aux premiers documents qu'il va visualiser. Ceux-ci auront un fort ancrage et vont construire sa représentation dominante du problème à résoudre. Un nouveau document différent de la représentation qu'il s'est forgée va introduire un phénomène de dissonance cognitive. La solution pour revenir à un état normal est de ne pas prendre en compte ce nouveau document. Ainsi, il sera très difficile pour un autre document qui n'est pas confirmatoire des précédents d'être retenu comme pertinent. L'internaute aura tendance à retrouver dans les documents futurs l'idée première qu'il s'est forgée sur les documents initiaux. Il aura ainsi des difficultés à abandonner sa première solution pour une autre. La rectification d'idées acquises est plus pénible pour un individu que l'apprentissage d'idées nouvelles pour lesquelles il ne possède pas encore de modèle. Aronson (1973) considère que la dissonance cognitive sera d'autant plus forte que l'effort aura été élevé pour acquérir la connaissance initiale.

On voit ainsi que l'ordre de lecture des résultats d'un moteur de recherche aura une influence sur ce qui sera retenu. Si l'internaute parcourt un ensemble de n documents dans un ordre différent, il n'aura pas la même connaissance au final.

b) Loi des petits nombres

Cette loi, présentée par Tversky et Kahneman (1974), Kahneman et al (1982) est une parodie de la loi des grands nombres utilisée en statistique. La « loi des grands nombres » repose sur un principe selon lequel des mesures conduites à partir d'échantillons représentatifs d'une population mère pourront ensuite être transposés à la population entière. Parler de « loi des petits nombres » correspond à l'idée selon laquelle les internautes ont du mal à apprécier la représentativité de l'information qu'ils exploitent et attribuent une confiance excessive à des conclusions issues de l'exploitation d'informations obtenues à partir d'échantillons non représentatifs. L'internaute a une mauvaise évaluation de la représentativité d'une information et peut par exemple considérer une information redondante ou une information fortement corrélée à la précédente comme confirmatoire d'une information initiale.

c) Clarté du document initial

La perception d'un document fait également intervenir la plus ou moins grande clarté des documents antérieurs. On peut considérer que la pertinence d'un document sera d'autant plus forte que les documents antérieurs sont clairs. En effet, plus la vision initiale est obscure, plus il faudra que le document courant soit clair, ceteris paribus, pour permettre à l'internaute de se forger une impression claire du problème qu'il a à résoudre.

Une autre voie peut être utilisée pour travailler sur le contexte de la requête. Elle repose sur la nature hypertextuelle du web. Il ne s'agit alors plus d'affiner la requête de l'internaute mais de reclasser les résultats en fonction du contexte. C'est l'objectif du *sensitive PageRank* (Haveliwala, 2003). Le PageRank calcule la pertinence d'une page en fonction de la pertinence des pages qui la citent : une page est importante si des pages importantes pointent vers elle. Le *sensitive PageRank* calcule la pertinence d'une page d'après la

pertinence des pages qui la citent et qui parlent du même sujet. Il s'agit donc d'un PageRank contextuel, une même page étant décrite par plusieurs PageRank contextuels.

Synthèse :

Une classification par facettes est un langage documentaire permettant de caractériser un document par plusieurs dimensions complémentaires. Ranganathan, le Père de la théorie des facettes, a défini cinq dimensions universelles pour caractériser un document. Ces dimensions sont regroupées sous l'acronyme PMSET (Personnalité, Matière, Espace, Energie, Temps). Dans les travaux de Ranganathan, la classification par facettes est avant tout un outil de classification de la connaissance.

Ce travail se revendique de la théorie des facettes et propose d'étendre l'usage de la classification par facettes à l'information disponible sur Internet. Toutefois, cette transposition au web suppose une certaine prise de distance par rapport au cadre défini par le père fondateur. Cette prise de distance est nécessaire pour des raisons liées à la nature de l'Offre et de la Demande d'information sur Internet.

L'information disponible sur Internet est bien différente de l'information disponible dans le rayonnage d'une bibliothèque. L'information contenue dans un article scientifique ou un ouvrage a été validée en amont par toute une série de filtres successifs (comité scientifiques, correcteurs…). Ce n'est que lorsque la publication scientifique aura franchi avec succès ces étapes successives qu'elle sera publiée. Sur Internet, il n'y a pas toujours de filtrage de l'information avant qu'elle ne soit publiée. Cela conduit à la nécessité d'un filtrage en aval qui est du ressort de l'internaute. Pour Maniez (1999), la facette décrit un « attribut fondamental d'un document ». Nous considérons que, dans le monde du web, la caractérisation de la validité d'une information est un attribut fondamental d'un document. Cette dimension est mal prise en compte par le concept de « pertinence » utilisé classiquement en « recherche d'information ».

Les profils d'usagers qui font de la recherche d'information sur le web sont très hétérogènes comparés aux profils des personnes qui font des recherches en bibliothèques. Le moteur de recherche est le creuset dans lequel vont se

retrouver des internautes aux attentes bien différentes. La théorie des facettes est puissante de ce point de vue puisqu'elle va permettre à l'internaute d'affiner son besoin. La théorie des facettes est alors plus utilisée comme outil de navigation de l'usager dans un océan documentaire que comme outil de classification de cet océan documentaire. Cette vision de la théorie des facettes orientée vers l'usager recouvre alors une partie des travaux qui s'inscrivent dans la logique de la pertinence centrée utilisateur.

CHAP 4 : CARACTERISATION DES FACETTES

Lorsqu'un internaute effectue une recherche d'information, il exprime son besoin en précisant de façon plus ou moins fine le sujet de sa recherche. Cette étape de spécification du sujet est décisive car elle va orienter la réponse qui lui sera renvoyée par le moteur de recherche. Nous considérons que cette expression du besoin peut être affinée par l'internaute à travers l'expression de dimensions complémentaires et orthogonales au sujet de la recherche. Nous proposons de décrire une page web par la tonalité de son discours, son degré de subjectivité, son niveau de lisibilité, son degré d'accessibilité, son niveau de centralité, sa fraîcheur et le trafic qu'elle génère. Dans la littérature, ces dimensions ont été observées de manière disjointe. Notre objectif est d'en présenter une vision fédérée.

Ce travail ne consiste pas à identifier, de façon exhaustive, toutes les facettes possibles qui correspondraient à tous les types de besoins. La démarche est ouverte : elle consiste à présenter certaines dimensions qui semblent utiles sans volonté d'exhaustivité. Nous nous sommes par exemple intéressés à des facettes qui correspondent à des besoins que peuvent avoir des professionnels de l'information et qui s'inscrivent dans une logique d'intelligence économique. Les besoins du professionnel de l'information seront eux-mêmes très différents selon qu'il s'agit de faire une recherche d'information dans le cadre d'une veille sectorielle (on pourrait privilégier dans ce cadre un indicateur de subjectivité) ou d'une veille d'image (la facette « tonalité du discours » apparaît adaptée puisqu'elle permet par exemple d'identifier des pages négatives dont le contenu est susceptible de porter atteinte à l'image de l'organisation). Nous souhaitons proposer une rapide typologie de ces facettes selon leur mode de validation.

Certaines facettes ont pour objectif de qualifier une page web en se rapprochant le plus possible d'une qualification humaine. Considérons par exemple la

facette qui détermine l'orientation positive ou négative d'une page web. Cette facette a pour objectif de proposer un découpage automatique entre pages positives et pages négatives d'un corpus qui se rapproche le plus du découpage qu'aurait réalisé l'internaute. De telles facettes se définissent par mimétisme ; il s'agit de se rapprocher le plus possible de la façon dont l'internaute pourrait appréhender la page web. La pertinence de ces facettes sera mesurée par la concordance entre la mesure automatique de l'indicateur et l'évaluation qui est faite de façon humaine par des juges.

Dans d'autres cas, l'internaute n'a pas les moyens d'apprécier la justesse de l'indicateur. Considérons par exemple le trafic d'un site web. Pour déterminer cet indicateur, il faut disposer d'une information statistique sur le nombre de visiteurs d'un site web. Cette information peut être estimée par un moteur de recherche qui a accès au fichier Log[70] mais elle est difficile à connaître par l'internaute. Dans ce cas, il n'est pas judicieux de soumettre le classement des pages suivant leur trafic pour le faire valider par l'internaute car il n'aurait pas les moyens de le faire. Les facettes de cette catégorie permettent à l'internaute d'obtenir des informations qu'il n'aurait pas pu obtenir (ou très difficilement) par analyse humaine.

Dans ce travail, nous n'avons pas décrit chacune des facettes avec le même niveau de détail. Certaines facettes sont développées, d'autres ne sont que citées sans faire l'objet d'investigations abouties.

Le *Tableau 18* classe les facettes selon leur niveau de validation et leur mode de développement dans ce travail.

[70] Le fichier Log est un fichier texte qui conserve la trace des différentes pages visitées par un internaute lors de sa navigation sur un site.

		Niveau de validation de la facette	
		Facette mimétique	Facette non mimétique
Niveau d'analyse en profondeur	Facette implémentée	▪ Polarité de la page web ▪ Niveau de subjectivité ▪ Niveau de lisibilité	▪ Niveau de centralité
	Facette non implémentée	▪ Page marchande ou non ▪ Page de contenu / page de lien	▪ Niveau de fraîcheur ▪ Folksonomie ▪ Distinction hub authority

Tableau 18 : les facettes selon leur niveau de validation et leur mode de développement

Pour chaque facette retenue, nous allons respecter la démarche suivante :

1. Dans un premier temps, nous présentons un état de l'art sur la question.
2. Nous nous intéressons ensuite à la transposition de cet indicateur sous forme de facette dans le contexte d'une information web. Ce paragraphe présente la chaîne de traitement déployée pour mettre en œuvre, de façon calculatoire, les algorithmes permettant de produire cette facette.
3. Le test de calibrage consiste à étalonner l'indicateur à l'aide d'un échantillon test. Cette étape permet d'affiner l'algorithme en jouant sur les pondérations.
4. La facette est alors testée sur un autre jeu de données en vue de mesurer la concordance entre jugements humains et évaluation automatique[71].
5. En dernier lieu une conclusion permet de dégager l'intérêt de la facette ainsi que ses limites et perspectives.

A. Valence et polarité de pages web

Il peut être intéressant, dans certains contextes, de savoir si une page web est orientée positivement ou négativement. Ceci est particulièrement utile dans le cas de veille d'image. L'idée est alors d'identifier le plus en amont possible d'éventuelles attaques informationnelles.

[71] Cette étape n'est mise en œuvre que dans le cas d'une facette qui doit être soumise à l'évaluation de juges.

Nous allons présenter, dans un premier temps, l'état de l'art sur le sujet puis préciser ensuite la façon dont nous avons implémenté cette facette dans notre travail.

1. Etat de l'Art de la notion de polarité

Les recherches relatives à l'orientation positive ou négative d'un texte sont contenues dans les travaux sur les émotions. Il existe six émotions (Ekman et Keltner-1997) valables quelles que soient les cultures : joie, colère, peur, tristesse, dégoût, surprise. La notion d'émotion est contextuelle, les émotions n'étant pas perçues de la même manière dans l'espace et dans le temps. Il semble donc difficile de mettre ce phénomène en équation en permettant à une machine de saisir les émotions dans toute leur finesse. Pour se rapprocher de cet objectif, il faut arriver à simplifier la réalité complexe en représentant les émotions à travers un modèle qui restreigne les dimensions possibles à un petit nombre plus facilement appréhendable. De nombreux travaux s'y sont employés. Ces travaux distinguent deux dimensions essentielles autour desquelles peuvent être décrites les émotions principales.

- la Valence d'un terme concerne l'orientation positive ou négative de ce terme : typiquement « bon » a une valence positive et « mal » une valence négative.
- l'Activation permet de savoir si le terme engage vers l'action ou l'inaction. Typiquement « ennui » a une activation faible là où « dynamique » a une valeur d'activation forte.

Ces deux dimensions se combinent pour donner lieu à une représentation sous forme de cercle (*Figure 27*) dans laquelle chaque terme est positionné sur ces deux axes (Cowie et al 2000).

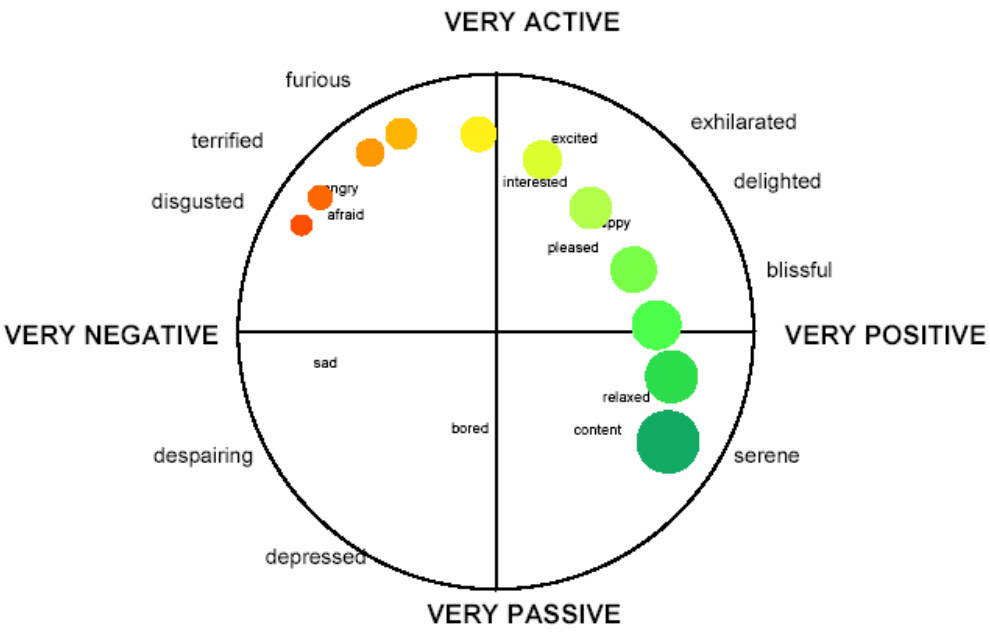


Figure 27 : le cercle des émotions d'après Cowie et al. (2000)

Whissel (1989) s'intéresse à la combinaison de ces deux dimensions et propose la *Figure 28* en retenant à chaque fois les 10% ou 25% des termes ayant la valence ou le degré d'activation le plus élevé ou le plus faible.

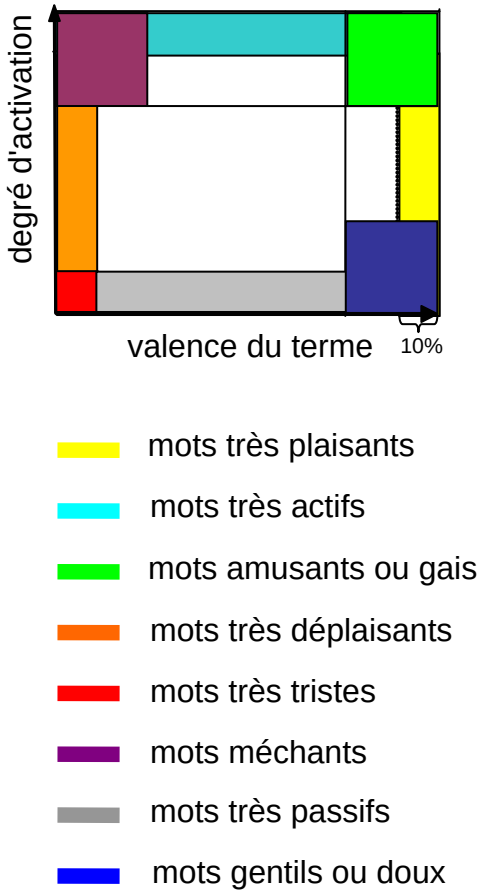


Figure 28 : matrice valence activation de Whissel

Dans ce travail, seule la dimension de valence nous intéressera.
L'analyse de la valence peut porter sur du texte ou sur un discours prononcé par un orateur. Dans le premier cas, ce sont des marqueurs textuels qui vont être utilisés pour déterminer la valence. Dans le second cas (Schroder -2004) le problème est plus complexe puisqu'il fait intervenir outre le texte prononcé, l'intonation du locuteur et les expressions du visage.

Nous allons nous attacher à voir de quelle manière il est possible de déterminer la valence d'un texte. Deux étapes successives doivent être franchies.
La première consiste à déterminer la polarité[72] des termes de la page web. On peut s'intéresser à certaines catégories de termes (adjectifs, pronoms, éléments de ponctuation), ou plus généralement s'intéresser à l'intégralité des termes. Selon les méthodes, la mesure de la valence d'un terme est qualitative ou quantitative. Dans le cas d'une mesure qualitative, le terme sera considéré

[72] La polarité désigne l'orientation positive ou négative d'un terme

comme positif, négatif ou neutre. Il n'est alors pas possible d'apporter de nuance et de granularité. Pour une plus grande finesse des résultats, on préfèrera une mesure quantitative. Dans ce cas, le terme prendra une valeur située sur un continuum.

Une fois que chaque terme du corpus a été qualifié, on est capable, dans la seconde étape, de déterminer par agrégation, la valence de la page web.

Ces deux étapes, présentées par Lacey (2005) sont reprises dans les deux paragraphes ci-dessous.

a) Déterminer l'orientation des termes d'un document

Notre objectif est de disposer d'une liste des termes français qualifiés par une mesure de leur polarité positive ou négative. Nous avons identifié 4 approches possibles pour parvenir à ce résultat.

(i) Utilisation de dictionnaires de valence existant

Un certain nombre de travaux antérieurs se sont attachés à qualifier la valence de termes, constituant ainsi des dictionnaires de valence à disposition du chercheur. Nous avons identifié deux sources d'information anglo-saxonnes sur le sujet.

La première est le General Inquirer (GI-2000) qui met à disposition 182 catégories de mots. Les termes sont répertoriés en positif, négatif ou neutre (case vide). Ces mots ont été classés à partir de la rubrique analyse de contenu du dictionnaire Harvard IV et de la rubrique analyse de contenu du dictionnaire Lasswell. Enfin 5 catégories sont issues des travaux de Semin and Fiedler (1988) en cognition sociale. Le *Tableau 19* fournit un extrait de 4 des 182 catégories du General Inquirer dans sa version de Novembre 2002.

Entry	Positive	Negative	Active	Passive
total	1915	2291	2045	911
A				
ABANDON		Negativ		
ABANDONMENT		Negativ		
ABATE		Negativ		Passive
ABATEMENT				
ABDICATE		Negativ		Passive
ABHOR		Negativ		Passive
ABIDE	Positiv		Active	
ABILITY	Positiv			
ABJECT		Negativ		Passive
ABLE	Positiv			
ABNORMAL		Negativ		
ABOARD				
ABOLISH		Negativ	Active	
ABOLITION				

Tableau 19 : 15 premières entrées du General Inquier

Il est à noter que ces listes sont discrètes, chaque terme ayant ou non l'orientation recherchée.

Le Whissell's dictionary of affect in langage (Whissell, 1989) fournit une liste de 8242 termes qualifiés de façon continue avec spécification pour chaque terme de sa valence, de son niveau d'activation et du caractère plus ou moins imagé du terme. Cette qualification a été réalisée manuellement par des étudiants volontaires, chaque terme ayant fait l'objet en moyenne de 8 évaluations. Le *Tableau 20* fournit les 15 premières entrées de ce dictionnaire. La colonne évaluation correspond à la valence du terme. Ce dictionnaire est intégré dans un outil capable de caractériser un document importé sur la base des trois dimensions analysées.

word	evaluation	activation	imagery
a	2	1.3846	1
abandon	1	2.375	2.4
abandoned	1.1429	2.1	3
abandonment	1	2	1.4
abated	1.6667	1.3333	1.2
abilities	2.5	2.1111	2.2
ability	2.5714	2.5	2.4
able	2.2	1.625	2
abnormal	1	2	2.4
aboard	1.8	1.875	2.8
abolition	1.5	2.1818	1.6
abortion	1	2.7273	2.6
about	1.7143	1.3	1.4
above	2.2	1.25	2.4
abroad	2.6	1.75	2.2

Tableau 20: 15 premières entrées du dictionnaire des affect compilé par Cynthia Whissel

Nous avons récupéré dans le cadre d'une coopération scientifique avec l'INRIA de Sophia Antipolis un dictionnaire de valence du français réalisé par le département de psychologie de l'université Catholique de Louvain (Belgique). Ce dictionnaire nous a été transmis par Robert Hogenraad. Il comporte 5744 entrées ce qui correspond à 35459 termes non lemmatisés de la langue française. Chaque terme est caractérisé par trois axes : actif / passif, émotionnel / non émotionnel et agréable / désagréable. Nous avons utilisé le dictionnaire comportant l'information relative à la dimension agréable / désagréable. Chaque terme a été qualifié par au moins 30 juges qui ont noté chacun le caractère désagréable ou agréable de chaque mot sur une échelle à 7 modalités (1 désignant un terme très désagréable et 7 un terme très agréable). La valence de chaque terme est définie par agrégation et prend la forme d'une mesure continue de valence comprise entre 15 (terme à valence très négative) et 70 (terme à valence très positive).

(ii) Utilisation d'un corpus qualifié humainement

Cette approche considère au départ un corpus de phrases. Pour chacune de ces phrases, plusieurs évaluateurs qualifient la valence de la phrase. Il existe un

certain nombre de corpus disponibles qui ont fait l'objet de cette évaluation humaine. Certains sont disponibles en ligne et peuvent servir à tester la pertinence de nouveaux algorithmes. On peut citer le cas du corpus de phrases relatives à des critiques de film utilisé par Pang et Al. (2002, 2004, 2005) et disponible en ligne[73]. Ce corpus est composé de 700 critiques de film positives et de 700 critiques de film négatives.

Une fois un tel corpus disponible, différentes solutions techniques sont possibles pour en inférer la valence des termes contenus dans ces phrases. Une approche retenue par Lacey (2005) revient à considérer que la valence d'un terme est la moyenne des valences des phrases dans lesquelles on retrouve ce terme. Des techniques d'analyses de données peuvent également être mobilisées. Elles sont généralement abondamment décrites dans la littérature au détriment souvent des résultats. Nous n'avons pas trouvé de corpus francophone de pages web dont les pages auraient fait l'objet d'une qualification humaine en terme de valence. Le corpus anglophone relatif aux critiques de films n'est pas généraliste. Ne disposant pas d'un corpus de textes francophones qualifiés humainement, il n'est pas envisageble, dans le cadre de ce travail, de recourir à cette méthode pour constituer un dictionnaire de valence française.

(iii) Utilisation de la cooccurrence de terme et propagation

Turney et Littman (2002), définissent l'orientation sémantique d'un mot (positif ou négatif) par la différence entre la force de son association avec des mots positifs d'une part et négatifs d'autre part. Les travaux de Hatzivassiloglou et Mckeown (1997) et Baroni et Vegnaduzzo (2004) ont montré que des termes qui avaient une valence voisine se trouvaient généralement associés au sein d'un même document voire au sein d'un même paragraphe. Ainsi est-il possible, à partir d'une base minimale de termes qualifiés humainement d'en inférer d'autres qui se retrouvent statistiquement dans le voisinage des premiers. Turney et Littman (2002) considèrent au départ quelques adjectifs positifs (good, nice, excellent, positive, fortunate, correct,

[73] http://www.cs.cornell.edu/people/pabo/movie-review-data/

superior) et quelques adjectifs négatifs (bad, nasty, poor, negative, unfortunate, wrong, inferior).

La polarité des autres adjectifs se définit donc de manière itérative. Les auteurs proposent une méthode astucieuse pour déterminer la propagation de la valence d'un terme à l'autre. Ils utilisent à cet effet le moteur de recherche Altavista. Considérons deux adjectifs A et B. A est un adjectif déjà polarisé et B appartient à la liste des adjectifs qui ne le sont pas. On peut connaître le nombre de pages web où les deux termes A et B sont présents conjointement séparés de part et d'autre par 10 mots maximum (cela correspond à la requête A NEAR B sous Altavista).

On peut connaître en utilisant un moteur de recherche le nombre de pages web dans lesquelles A est présent (noté n_A), B est présent (noté n_B) et enfin A et B sont présents (noté $n_{A\&B}$)

En effectuant le ratio $2n_{A\&B}/(n_A + n_B)$, on obtient un indicateur compris entre 0 et 1. Si cet indicateur vaut 0, cela signifie que les adjectifs ne sont jamais associés. Si cet indicateur vaut 1, cela signifie que ces deux adjectifs apparaissent toujours ensemble. Plus la valeur de l'indicateur est forte, plus les adjectifs sont souvent présents l'un avec l'autre. On émet alors l'hypothèse qu'ils ont la même polarité. Il est à noter que la commande NEAR n'est plus maintenue par Altavista. On peut donc rechercher la présence de deux mots dans une même page web sans descendre à un niveau de granularité plus fin comme c'était le cas avec la commande NEAR. L'intérêt de cette approche est de faire abstraction d'une thématique spécifique puisque le corpus de référence est le web dans son intégralité. On évite ainsi tous les problèmes d'échantillonnages et de représentativité.

(iv) Utilisation de dictionnaire de synonymes

La revue de la littérature nous a permis de trouver une dernière méthode permettant de définir l'orientation positive ou négative d'un terme. L'hypothèse de cette méthode est que deux synonymes auront des valences équivalentes. Cette méthode, comme la précédente, consiste à partir d'un petit ensemble de termes dont on connaît la valence et d'en déduire de proche en proche la valence des termes qui leur sont synonymes. Cette méthode est très

rapide. Elle est compatible avec l'exploitation de dictionnaires en ligne dont certains sont disponibles pour la langue française. Cette méthode a été mise en œuvre de façon différente par Kamps et al. (2004) et Esuli et Sebastiani. (2005).

Kamps et al. (2004) raisonnent à partir du dictionnaire WordNet. Ils partent de deux mots polaires : good et bad. La distance entre deux termes s'apprécie par le chemin de longueur minimale entre le terme recherché et le terme *good* ou *bad*. On ne peut qualifier avec cette méthode que les adjectifs qui ont un lien de synonymie direct ou indirect avec good ou bad.

Esuli et Sebastiani (2005) ont élargi l'ensemble de départ qualifié positivement ou négativement en considérant la liste de Turney et Littman (2002) et non seulement les termes good et bad.

b) Déterminer l'orientation d'une page web

Une fois la valence des termes d'une page connue, il est possible d'agréger les résultats pour obtenir la valence de cette page. Toutefois, pour être rigoureux, cette agrégation doit être précédée d'un traitement linguistique et lexical du corpus.

(i) Prise en compte des règles linguistiques

Si un terme, ayant une valence positive, est par exemple précédé d'une négation, alors la négation confèrera à la phrase une valence négative. Il faut donc, pour conduire une analyse plus fine, être capable d'identifier certains éléments linguistiques. Deux peuvent être mentionnés :

- Le premier concerne la négation ou certaines formes verbales. Si un terme se retrouve entre la négation et une marque de ponctuation, alors le terme prend une valence inverse. Le *Tableau 21* comporte les principales formes négatives utilisées par Mukras (2004)

didnt	didn't	arent	aren't
dont	don't	aint	ain't
isnt	isn't	havent	haven't
wont	won't	couldnt	couldn't
cant	can't	wasnt	wasn't
wouldnt	wouldn't	shouldnt	shouldn't
werent	weren't	not	

Tableau 21 : formes négatives utilisées par Mukras

Certaines unités lexicales composées ont une valence différente des valences des termes qui la composent (Bestgen et al., 2004). L'expression « avoir mangé son pain blanc » a une valence plutôt négative alors qu'aucun de ces termes n'a de valence négative.

- Le second concerne certains adverbes ou termes qui peuvent avoir un rôle de renforcement ou d'atténuation. Ainsi par exemple « très gentil » a une valence plus forte que gentil. Si on considère un indicateur de valence continu, il est important de regarder le voisinage du terme pour voir si son sens est renforcé ou non, atténué ou non. Cette approche est développée par Casey Whitlaw et al. (2005) qui parlent *d'appraisal group* pour caractériser ces expressions. Chacune d'elles peut être définie par 4 dimensions :
    - l'attitude se définit autour des émotions personnelles (joie, tristesse) ou l'évaluation d'entités extérieures (héroïque, idiot)
    - l'orientation (positive ou négative)
    - la force permet de mesurer l'intensité de l'orientation
    - la polarité est utilisée par l'auteur pour caractériser la présence d'un terme qui annule la valence du terme principal.

Le *Tableau 22* fournit trois exemples considérés par les auteurs.

"heureux"		"très heureux"		"pas très heureux"	
Attitude	Affect	Attitude	Affect	Attitude	Affect
orientation	positive	orientation	positive	orientation	négative
Force	neutre	Force	forte	Force	faible
polarité	non marqué	polarité	non marqué	polarité	non marqué

Tableau 22 : exemple de appraisal group d'après Whitlaw(2005)

(ii) Prise en compte des unités lexicales

Une fois ces traitements réalisés, il est possible de déterminer la valence générale de la page. Il est important de préciser que la plupart des études expérimentales sur la valence ont porté sur des textes courts de type intervention sur des forums de discussion.

D'autre part, alors que nous cherchons à définir la valence moyenne de la page, d'autres recherches portent sur l'identification de zones d'un texte ayant une valence particulière. En effet selon Yi (2003), il peut être intéressant de descendre à un niveau plus fin pour détecter la valence de sous composantes de la page. Les clients d'un constructeur automobile peuvent être globalement satisfaits de leur véhicule mais cela ne les empêche pas d'être critiques sur un point. Si on se contente d'une analyse générale, la critique passera inaperçue.

La valence d'un document est obtenue par combinaison des valences des termes qui la composent par exemple par la somme des valences des termes positifs moins la somme des valences des termes négatifs. La dispersion de la valence des termes est également intéressante. Une dispersion forte dans les valences des termes d'une page correspond à une certaine confusion des sentiments (Cowie-2003). Pour pouvoir caractériser la valence d'une page web, Whissel propose de comparer la valeur obtenue pour une page web avec la valence générale calculée sur tout le corpus. L'indicateur de valence est en moyenne de 1,84 avec un écart type de 0,44.

L'approche retenue par le General Inquirer propose une comparaison des valences des pages web du corpus entre elles plutôt qu'une comparaison avec une valence moyenne valable pour la langue anglaise quel que soit le contexte.

Une fois la valence définie, les résultats sont affichés de façon plus ou moins visuelle. *Schröder* (2004) propose une visualisation sur le cercle des émotions (*Figure 29*).

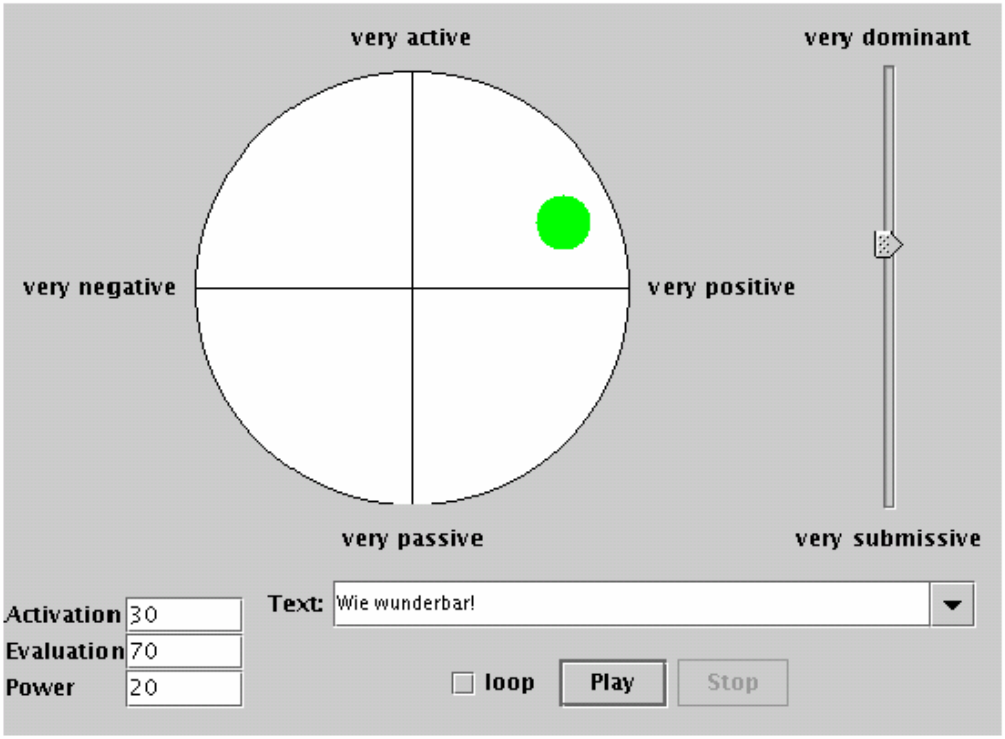


Figure 29 : l'interface de Schröder est disponible en ligne : http://mary.dfki.de/emotional.html

Après avoir défini les pages positives et négatives d'un corpus de pages web, il est intéressant de regarder la structure relationnelle d'un tel corpus. Les pages objectives pointent-elles plus volontiers vers des pages objectives ? Si tel était le cas, la structure hypertextuelle du corpus pourrait servir de source d'information complémentaire pour apprécier la subjectivité d'une page dans son contexte.

2. La polarité comme facette : protocole et mise en œuvre calculatoire

Notre objectif est de construire un indicateur qui permette de mesurer la valence d'une page web. L'indicateur de valence d'une page web que nous avons retenu correspond à la moyenne des valences des termes constituant le texte de la page à qualifier. Nous avons à notre disposition un dictionnaire de valence qui, pour 35459 termes non lemmatisés de la langue française, renvoie une valeur comprise entre 15 et 70. Plus le terme a une valeur faible, plus sa valence est négative. Plusieurs précisions doivent être apportées pour comprendre les choix qui ont été faits :

- Il est à noter que dans le dictionnaire primitif, figure un certain nombre de termes qui possèdent plusieurs valeurs de valence. Etant incapable de mettre en œuvre une logique d'analyse sémantique dans le cadre de cette étude, nous avons choisi d'éliminer tous les termes à sens multiple. Par exemple le mot « bête » a une valence de 21 et de 40 selon la signification (adjectif stupide ou nom commun synonyme d'animal) donnée à ce terme. 87 termes ont été ainsi enlevés du dictionnaire pour retenir les termes non ambigus.
- A noter également que nous ne prenons pas en compte la négation qui est de nature à inverser la valence du terme sur lequel elle porte. Nous avons considéré que l'emploi d'un vocabulaire négatif n'est pas neutre et que si l'auteur fait le choix d'une tournure négative avec un mot négatif, il aurait pu choisir d'utiliser une tournure positive ce qui n'aurait pas changé le sens de la phrase mais ce qui aurait donné à la phrase une tournure plus positive.
- En établissant la moyenne de la valence des termes constituant une page web, on fait l'hypothèse que les pages ont une orientation soit positive, soit négative. Il est clair que l'usage de la moyenne est de nature à rendre neutre des pages positives sur certains paragraphes et négatives sur d'autres. L'utilisation de l'écart type pourrait palier ce problème.

Notre indicateur automatique de valence est le fruit d'un processus de traitement semi automatique mettant en œuvre divers logiciels :

- Le logiciel web pipe pro[74] pour aspirer le contenu d'un ensemble de page web avec un niveau de profondeur donné
- le logiciel text pipe pro[72] pour extraire un à un les mots d'un corpus de page web.
- Le logiciel Microsft Access pour établir une relation entre le corpus préformaté par le logiciel Text Pipe Pro et un dictionnaire de valence.

Nous renvoyons le lecteur intéressé par l'étude de la complexité temporelle des algorithmes et des temps de calcul à l'annexe 16.

[74] www.datamystic.com/textpipe.html

Cet indicateur automatique a été obtenu à l'issue d'un processus itératif d'affinage successif et de confrontation aux usagers. L'analyse de concordance entre indicateur automatique et jugements humains a été réalisée sur divers corpus successifs. Cette confrontation au terrain a permis de faire évoluer l'indicateur en jouant sur certains paramètres. L'indicateur a été soumis à un premier corpus de pages ce qui a permis d'affiner et de calibrer le modèle. Une fois le modèle calibré, il a été soumis à un autre jeu de pages web pour voir comment il se comportait.

3. Calibrage

La puissance de cet indicateur s'apprécie par la concordance entre la valeur calculée par l'algorithme et le jugement d'usagers sur un jeu de pages web : une page qualifiée par l'usager de négative, de positive ou de neutre doit être idéalement estimée comme telle par l'indicateur calculé. Le problème est que tous les juges n'ont pas forcément la même appréciation sur une page et qu'on observe une variabilité interindividuelle. Pour cette raison, nous avons fait le choix de nous limiter à un corpus de pages consensuelles pour lesquelles les usagers avaient exprimé un choix voisin pour le critère considéré. Notre travail consiste donc dans un premier temps à définir un corpus de pages consensuelles puis à comparer, dans un second temps, l'avis des juges avec l'indicateur automatique.

a) Comment trouver des pages consensuelles.

Lorsqu'on demande à des internautes de s'exprimer sur le caractère positif ou négatif d'une page web, ceux-ci affichent des positions qui peuvent être différentes. Difficile dans un tel cas de mesurer l'efficacité de notre indicateur automatique de valence. Nous avons donc choisi de partir d'un corpus de pages web consensuelles pour lesquelles les internautes exprimaient des opinions voisines. Ce corpus a été déterminé de la façon suivante :

- Nous considérons au départ une liste de 75 pages web. Ces pages web constituent un échantillon bien trop faible pour pouvoir être considéré comme représentatif du web. Les 75 pages web correspondent à 5 pages web pour chacune des 15 requêtes suivantes : Dalai lama ; fumer ;

polygamie ; Darfour ; éléphant ; dinosaure ; effet de serre ; synchronicité ; petit Prince ; Caméléon ; nucléaire ; Total ; Sarkozy ; Goji ; Régime[75].

- Notre indicateur de valence étant construit essentiellement à partir de critères textuels, le texte de chacune de ces 75 pages web est aspiré par le logiciel Text pipe pro[76]. On observe à l'usage que la technologie de Text pipe pro n'était pas capable de récupérer proprement le texte de certaines pages web, les caractères accentués de ces pages se transformant en caractères inexploitables. Ces pages ont été supprimées du jeu de données pour obtenir un corpus réduit de 52 pages. Ce jeu est fourni en annexe 3.
- Ce corpus de 52 pages web a été soumis à 75 étudiants inscrits en seconde année de Techniques de Commercialisation de l'IUT de Toulon en janvier 2008. Les étudiants étaient installés en binôme devant un ordinateur. Ils devaient visualiser au moins 10 pages de leur choix parmi les 52 pages et qualifier le niveau de valence de ces pages sur une échelle à 5 modalités: page très négative, page plutôt négative, page neutre, page plutôt positive, page très positive. Les étudiants sont deux par machine ; il leur est demandé de se mettre d'accord sur une des modalités proposées pour la page considérée. Si ce n'est pas possible, alors ils cochent les deux modalités. Cette qualification humaine n'est pas aisée car les étudiants doivent juger de la valence de la page dans l'absolu. Il leur est difficile de disposer de repères pour évaluer des pages différentes. Il a été pour cette raison nécessaire de reformuler la consigne de notation aux évaluateurs en prenant des exemples extrêmes. Une page sera considérée comme négative si elle exprime des faits ou des opinions hostiles, si le ton utilisé dans l'écriture est négatif, si cette page, par son environnement typographique ou colorimétrique, donne l'impression au juge d'être critique, violente. A l'inverse une page positive va utiliser des termes emphatiques pour décrire un phénomène. Elle sera orientée positivement.
- Les étudiants ont la liberté de choisir 10 pages de leur choix parmi les 52. Certaines pages web sont donc plus analysées que d'autres. Nous avons

[75] Nous avons considéré un ensemble de requêtes couvrant différents domaines avec pour chacune d'elles un petit nombre de réponses. Nous ne sommes donc pas dans les conditions d'utilisation réelles de notre algorithme qui est plutôt testé sur un grand nombre de pages d'une même requête.

[76] www.datamystic.com/**textpipe**.html

retenu, pour la suite de l'analyse, toutes les pages qui avaient été qualifiées par au moins 6 juges (6 groupes de deux juges devrait-on dire). Cette contrainte nous a conduits à écarter 14 pages web et à nous limiter à un corpus de 38 pages. Considérons par exemple *Tableau 23* l'évaluation d'une page du corpus.

page	très négative	plutôt nég	neutre	plutôt positive	très positive	total
www.tabac-cigarette.com\default.htm	24	8	1	2	0	35

Tableau 23 : exemple de qualification d'une page du corpus

Cette page a été qualifiée de très négative par 24 binômes sur 35.

- Pour chaque page web, nous supprimons systématiquement deux jugements supposés aberrants correspondants aux modalités extrêmes de l'échelle. Ainsi les modalités de la page web retenues pour la suite de l'analyse dans l'exemple précédent sont présentées *Tableau 24*.

page	très négative	plutôt nég	neutre	plutôt positive	très positive	total
www.tabac-cigarette.com\default.htm	23	8	1	1	0	33

Tableau 24 : qualification d'une page du corpus après suppression des valeurs aberrantes

- L'échelle à 5 composantes est ensuite réduite à une échelle à 3 composantes, les modalités « très négative » et « plutôt négative » sont regroupées en une modalité « négative ». Idem pour les deux modalités positives. L'évaluation de la page précédente est présentée *Tableau 25*.

page	négative	neutre	positive	total
www.tabac-cigarette.com\default.htm	31	1	1	33

Tableau 25 : qualification d'une page du corpus après regroupement des modalités

- Dans le cas d'une page consensuelle, on observe des réponses convergentes autour des mêmes modalités. C'est le cas de la page qui nous intéresse qui apparaît comme une page négative. Dans d'autres cas, la forte dispersion des jugements humains conduit à considérer que la page n'est pas consensuelle. Si l'on considère que l'écart entre les 5 modalités (très négatif, plutôt négatif, neutre, plutôt positif, très positif) est le même, on peut se ramener à une note entre 1 et 5. On pourra alors considérer comme consensuelles, les pages dont l'écart type des notes est faible. Pour déterminer les pages consensuelles, nous avons combiné ce critère automatique qui repose sur l'écart type avec le jugement humain. Sur les 38 pages de notre corpus, nous observons la distribution des réponses représentée *Figure 30*.

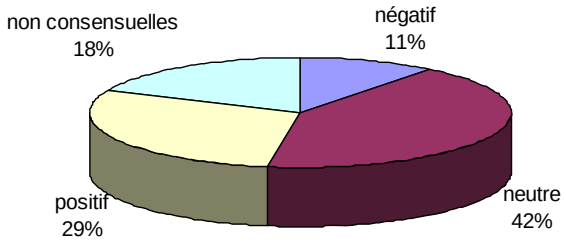


Figure 30 : Valence des pages définie par les experts humains

31 pages (soit 82% des pages étudiées) sont donc considérées comme consensuelles. C'est sur ces pages que l'analyse va être poursuivie. Cet échantillon de 31 jugements peut sembler insuffisant au regard de la statistique traditionnelle. En fait, étant donné que nous avons trois modalités, la taille de l'échantillon doit être supérieure à 18 selon Cicchetti et Fleiss (1977) et de 25 d'après Fermanian (1984), Les conditions minimales de représentativité sont donc respectées.

Le tableau de l'annexe 4 fournit pour chaque page son statut au regard des jugements exprimés par les juges humains.

Dans ce travail, nous avons confronté le jugement qualitatif des juges avec notre indicateur calculé qui est quantitatif. Il s'est agi de trouver des valeurs de coupes « idéales », de sorte à ce qu'en dessous de telle valeur on considère la

modalité « valence négative » au dessus de telle autre « valence positive » et entre les deux « valence neutre ». Ce calibrage a été réalisé grâce à nos 31 pages. Nous avons considéré que les pages de valence moyenne inférieures à 41,45 étaient de valence négatives, que celles dont la valence était supérieure à 44,62 étaient de valence positive et que les autres étaient de valence neutre.

Lorsqu'on confronte les indicateurs automatiques et les indicateurs issus des jugements humains sur les pages consensuelles, on obtient le *Tableau 26*.

		Jugements humains sur pages consensuelles		
		Positif	Neutre	négatif
Jugements automatiques	Positif	5	3	1
	Neutre	6	11	0
	Négatif	0	2	3

Tableau 26 : Indicateur de valence : confrontation de l'indicateur automatique et des jugements humains

Les cases grisées correspondent à la concordance entre l'évaluation humaine et l'évaluation automatique. Cette concordance a donc lieu dans 19 cas sur 31 soit 61%. Pour la valence positive, on observe une précision de 5/9 soit 55% et un rappel de 5/11 soit 45%.

Sur la base de ce premier tableau, plusieurs pistes d'amélioration ont été envisagées, l'objectif étant de se rapprocher d'un indicateur pour lequel la concordance entre les valeurs automatiques et le jugement humain serait parfaite. Nous allons présenter deux de ces perfectionnements.

- Le premier perfectionnement a consisté à retenir dans le dictionnaire exclusivement les termes à valence positive ou négative. On a ainsi conservé les termes du dictionnaire à valence inférieure à 30 (termes négatifs) et à valence supérieure à 45 (termes positifs). Le fait de se restreindre aux termes à valence positive et négative élimine les termes neutres qui par leur nombre peuvent diluer l'indicateur et le tirer vers un indicateur moyen pour toutes les pages. Ce choix nous a permis de gagner un peu de pertinence. Nous avons considéré que les pages de valence

moyenne inférieures à 44,67 étaient de valence négative, que celles dont la valence était supérieure à 50,14 étaient de valence positive et que les autres étaient de valence neutre. Lorsqu'on confronte les indicateurs automatiques et les indicateurs issus des jugements humains sur les pages consensuelles, on obtient le *Tableau 27*.

		Jugements humains sur pages consensuelles		
		Positif	Neutre	négatif
Jugements automatiques	Positif	3	1	0
	Neutre	7	14	1
	Négatif	1	1	3

Tableau 27 : *Indicateur de valence : indicateur automatique versus jugements humains perfectionnement 1*

Les cases grisées correspondent à la concordance entre l'évaluation humaine et l'évaluation automatique. Cette concordance a donc lieu dans 20 cas sur 31 soit 64,45%. Pour la valence positive, on observe une précision de 75% et un rappel de 27%

- La seconde piste d'amélioration a consisté à s'intéresser au nombre de mots de la page web étudiée retrouvés dans le dictionnaire. On a pu observer que plus la page comporte un grand nombre de mots dans le dictionnaire de valence et plus l'indicateur automatique est concordant avec les juges humains. Le *Tableau 28* compare la taille minimale de la page retenue en nombre de mot et le taux de concordance c'est-à-dire le rapport entre le nombre de pages où il y a concordance entre évaluateur humain et algorithme automatique et le nombre total de pages qualifiées.

Nombre de mots minimum à considérer pour que la page soit prise en compte	Taux de concordance observé sur l'échantillon de calibrage entre jugements humains et automatiques
0	64,45%
100	65%
200	77%
300	91%
400	100%

Tableau 28 : concordance observée entre jugements humains et automatiques

Si on conserve parmi les 31 pages web uniquement celles qui ont au moins 100 mots qui se retrouvent dans le dictionnaire, alors on a une concordance de 65% entre jugements humains et jugements automatiques.

On peut observer sur ce tableau que la concordance augmente avec la taille de la page web. Lorsqu'on réduit le corpus aux pages qui ont plus de 400 mots, la concordance est parfaite entre juges humains et algorithme automatique. Ces chiffres intéressants masquent le fait que ces résultats positifs ont été observés non plus sur les 31 pages mais seulement sur les 7 pages qui ont plus de 400 mots dans le dictionnaire.

A l'issue de cette démarche de calibrage, on arrive à un indicateur de valence calculé comme la moyenne des valences des termes de la page dont la valence est inférieure à 30 (termes négatifs) ou supérieure à 45 (termes positifs). Cet indicateur s'affine au fur et à mesure que le nombre de mots reconnus dans le dictionnaire de valence augmente. La valeur optimum est obtenue pour les pages de plus de 400 termes reconnus mais les résultats ne sont plus représentatifs en raison de la taille de l'échantillon (7 observations).

Il s'agira donc de constituer un nouvel échantillon de sorte qu'il soit possible de retenir des pages web qui ont plus de 400 termes dans le dictionnaire. Si ce résultat initial venait à être confirmé, cela voudrait dire qu'il serait possible de définir l'orientation positive ou négative d'un texte dans certains cas maîtrisés.

4. Test

Pour le test, nous avons considéré au départ un corpus de 63 pages web (annexe 5). Pour chacune de ces pages web, on observe que 400 mots au moins figurent dans le dictionnaire de valence. Ces 63 pages web sont soumises aux juges dans les mêmes conditions que celles de l'expérimentation précédente. Les résultats sont fournis en annexe 6. Il est à noter que cette expérimentation a pour objectif de valider l'indicateur de valence et de calibrer l'indicateur de subjectivité de page web. Cette expérimentation simultanée permettra d'étudier les relations entre ces deux indicateurs. Le tableau de l'annexe 6 présente les résultats de cette expérimentation. Pour 22 des 63 pages, nous n'avons pas obtenu un nombre de jugements humains suffisants pour poursuivre l'analyse (cases notées nbinsuff dans la case correspondante du tableau de l'annexe 6). Notre corpus a donc été réduit à 41 pages web puis restreint encore par la prise en compte des pages consensuelles. Nous avons utilisé la même logique que précédemment pour déterminer le caractère consensuel ou pas de chacune des pages du corpus. Les résultats observés pour ces 41 pages web sont présentés *Tableau 29*.

Type de page	nombre
Pages non consensuelles	12
Pages négatives	9
Pages positives	10
Pages neutres	10
Total	41

Tableau 29 : Valence des pages définies par les experts humains

On observe donc que dans 71 % des cas, les juges sont d'accord sur la nature de la page qu'ils ont à qualifier.

Le *Tableau 30* permet de confronter le calcul automatique et le calcul issu des juges pour ces 29 pages consensuelles. Il est à noter que ce corpus de 29 pages se caractérise comme le précédant par le fait que certaines pages ne sont pas récupérables par le logiciel Text pipe pro. Ces pages ont été récupérées manuellement pour ne pas appauvrir la taille de notre échantillon.

		Jugements humains sur pages consensuelles		
		Positif	Neutre	négatif
Jugements automatiques	Positif	3	0	1
	Neutre	6	10	3
	Négatif	0	0	6

Tableau 30 : *Indicateur de valence : confrontation de l'indicateur automatique et des jugements humains*

Pour la valence positive, on observe une précision de 75% et un rappel de 33%. Notre étude repose au final sur 29 observations ce qui est suffisant pour effectuer le test de Kappa.

Dans 19 cas sur 29, soit 65% des cas, il est possible de déterminer, par l'algorithme automatique, un statut de page web qui soit concordant avec le jugement humain. La concordance aléatoire est de 34%. Le test de Kappa, présenté dans la méthodologie statistique, appliqué sur ce tableau donne un résultat de 46%. L'accord est modéré au sens de Landis et Koch (1977) mais il faut tenir compte du fait que les modalités présentent un déséquilibre symétrique (Briand et al, 2002).

Le Tableau 31 précise les valences des pages non consensuelles. On observe que l'indicateur de valence balaie tout le spectre des valences possibles (valence négatives à positives) sans survalorisation particulière des valences moyennes.

Page	valence
www.lesafriques.com\international\la-crise-des-subprimes-epargne-l-afrique.html@Itemid=5	38
www.wsws.org\francais\News\2002\avril02\26avril02_boycottfr.shtml	42
pofpom.skynetblogs.be\post\3360338\programme-du-mois--5-fruits-et-legumes-par-jo	45
www.infosdelaplanete.org\2867\l-assassinat-de-la-voiture-electrique.html	45
culturepolitiquearabe.blogspot.com\2007\12\my-life-arab-movie-boycott-et.html	47
www.20minutes.fr\article\202149\Monde-La-voiture-electrique-passee-au-crible.php	48
www.amazon.fr\m_25E9thode-fruits-l_25E9gumes-par-jour\dp\2702904602	49
www.village-justice.com\forum\viewtopic.php@p=138279	50
fr.wikipedia.org\wiki\Facebook	50
www.presse-citron.net\@2008_2F01_2F05_2F2912-facebook-petites-betises-et-consequences-inattendues	51
www.fluctuat.net\6006-Facebook-vs-Meetic	51

Tableau 31 : Tableau exprimant la valence des pages non consensuelles

5. Discussion des résultats et conclusion

Il convient tout d'abord de préciser que l'indicateur de valence correspond à une réalité partagée par des internautes non professionnels. En effet on a pu observer aussi bien dans le test de calibrage que dans le test définitif que le pourcentage de pages consensuelles était compris entre 70% et 80%.

Le jugement humain qui va permettre de qualifier une page en lui affectant une valence positive ou négative fait entrer en ligne de compte plusieurs éléments comme le recours à des éléments textuels négatifs ou positifs, l'interface graphique, le choix des images et des couleurs, l'url. Il est difficile de rendre compte de cette complexité et de la traduire par un indicateur automatique. L'indicateur de valence que nous avons retenu est lexical : il apprécie la valence d'une page web par la moyenne des valences des termes constituant le texte de la page à qualifier. Afin d'affiner l'algorithme, nous avons restreint l'analyse à des pages web présentant au moins 400 mots dans le dictionnaire de valence. Le test de concordance a permis de tester l'efficacité de la méthode en confrontant pour chaque page web le score attribué par l'indicateur automatique et le score humain lorsqu'il fait l'objet d'une mesure consensuelle. Les résultats obtenus sur le test de concordance sont modérés, la concordance s'opérant dans 65% des cas.

B. Degré de subjectivité d'une page web

1. Etat de l'Art de la notion de subjectivité

Dans cette partie, nous allons nous intéresser à la mesure du degré de subjectivité d'une page web. L'objectivité d'un document est définie par Wiebe (2000) comme la capacité d'un document à présenter une information factuelle par opposition à une information subjective qui exprime une opinion ou une évaluation. Ce continuum qui irait d'une information objective à une information subjective s'avère pertinent dans un processus d'intelligence économique. En fonction de sa problématique, le veilleur peut privilégier une information subjective ou objective. Une information subjective sera recherchée par le veilleur lorsque celui-ci effectue une veille d'image où qu'il cherche à identifier les menaces qui pèsent sur son activité. Au contraire

lorsque le veilleur doit réaliser un état de l'art sur un sujet, il privilégiera une information qui écartera l'information subjective.

Nous proposons d'identifier le degré de subjectivité d'une page web de façon automatique. Cette approche constitue une facette d'un outil de veille permettant au veilleur, au moment de sa requête, de pouvoir choisir, outre le thème de sa recherche, une information possédant un niveau de subjectivité qu'il spécifierait.

Lorsqu'on souhaite déterminer la subjectivité d'un texte, il faut trouver dans ce texte des marqueurs caractéristiques de la subjectivité. Plusieurs approches ont pu être proposées pour identifier ce type de marqueurs. Ces marqueurs identifiés dans la littérature sont surtout de nature stylométrique. Le style d'un document décrit les dispositifs structurels de l'écriture : ponctuation, taille des mots, forme de phrase, utilisation de certains mots. Nous allons en présenter quelques-uns.

Beaudouin et al. (2003) ont mis en évidence le rôle des pronoms personnels. La présence du « je » « j' », « m' », « moi » dans un texte correspond à l'expression d'une information engagée. Beaudouin observe que ces pronoms sont plus présents dans les pages personnelles et moins dans les pages marchandes qui recourent plus volontiers au « vous ».

Stammatatos (2000) a montré que certains éléments de ponctuation pouvaient s'avérer discriminants pour déterminer le genre d'une page web. Le point d'exclamation apparaît de ce point de vue comme un marqueur du degré de subjectivité d'un texte.

Dans les deux travaux que nous venons de citer, les marqueurs observés sont des termes très fréquents dans le corpus. Paradoxalement ces marqueurs correspondent à ce qui est convenu d'appeler le bruit statistique.

Wiebe (2000) a travaillé sur une relation qui associerait le caractère subjectif d'un texte à la présence d'adjectifs dans ce texte. En effet, les adjectifs seraient porteurs, plus que les autres termes, de l'émotion. Wiebe fait référence à des travaux antérieurs qui montrent que le nombre moyen d'adjectifs par phrase est un indicateur de la subjectivité d'un texte. Ainsi Wiebe a testé cette hypothèse à partir d'un corpus de 486 références subjectives et 515 références objectives (pré-qualifiées par des experts). Il arrive à la conclusion que la probabilité pour qu'une phrase soit subjective sachant qu'elle contient un adjectif est de 55,8%

alors que la probabilité qu'une page du corpus soit subjective est de 48%. Cette approche brute est d'une mise en œuvre calculatoire rapide mais manque de précision. Elle a ensuite été affinée par Wiebe qui distingue trois formes de subjectivité :

- l'évaluation positive que l'on retrouve dans des termes comme « fascinant »
- l'évaluation négative que l'on retrouve dans des termes comme « terrible »
- la spéculation que l'on retrouve dans des termes comme « probablement ».

Wiebe expose que les adjectifs ne sont pas tous polarisés négativement ou positivement. Certains sont neutres et ne transmettent pas de subjectivité (« domestique », « médical », « rouge »). D'autre part, Wiebe distingue la subjectivité absolue de la subjectivité relative. La subjectivité absolue est valable quel que soit le contexte (horrible est un adjectif qui introduit une expression subjective dans tous les contextes). La subjectivité relative est beaucoup plus dépendante d'un domaine. Certains adjectifs peuvent avoir une connotation négative dans certains domaines et pas dans d'autres.

Dans la littérature francophone sur le sujet, on peut citer le travail de Trubert-Ouvrard (2001) sur l'antéposition ou la postposition de l'adjectif par rapport au nom. L'auteur étudie un certain nombre de discours d'hommes politiques français à l'occasion d'une élection présidentielle. Ils considèrent que lorsque l'adjectif suit le nom cela traduit une subjectivité alors que celle-ci est absente lorsque l'adjectif précède le nom. Ainsi comparons l'expression « c'est la raison principale » et « c'est la principale raison ». Dans le premier cas qui illustre la postposition de l'adjectif, on a l'impression que l'énonciateur se porte garant. « C'est la raison principale » est équivalent à « c'est la raison que je considère personnellement comme principale ». En utilisant l'antéposition « c'est la principale raison », on permet à la raison d'être inscrite dans la réalité passée. On se retranche derrière une raison présentée comme déjà existante. La postposition de l'adjectif est donc perçue comme plus subjective que l'antéposition.

On s'aperçoit, en lisant les perfectionnements de Wiebe, que la notion de subjectivité d'un texte est très liée à celle d'orientation positive ou négative d'un texte. Un texte sera subjectif lorsqu'il aura une polarité positive ou négative. Il sera objectif lorsqu'il aura une polarité neutre.

2. La subjectivité comme facette : protocole et mise en œuvre calculatoire

L'étude préalable de la littérature nous conduit à identifier différents éléments susceptibles d'entrer en jeu dans la détermination d'un indicateur automatique de mesure de la subjectivité d'une page web. Ces éléments n'ont pas tous été validés dans le cas de pages web ni dans le cas de la langue française mais on peut penser qu'il y a là un socle permettant de définir un indicateur de mesure automatique de la subjectivité d'une page web francophone. Ces éléments sont listés ci-dessous et font l'objet de sous hypothèses qui sont testées :

1. Une page serait d'autant plus subjective qu'elle posséderait un grand nombre d'adjectifs.
2. L'emploi d'un verbe au conditionnel prédisposerait une page à avoir plutôt une orientation subjective.
3. L'emploi de la première personne du singulier (je, mon, mes, miens, moi, j') prédisposerait davantage à une page subjective.
4. L'emploi de termes à polarité extrême (termes très positifs ou très négatifs) prédisposerait davantage une page à être subjective.
5. Certains éléments de ponctuation (!) se retrouvent volontiers dans une page subjective.
6. L'adresse url d'une page web comporte des informations qui peuvent dans certains cas orienter vers une page objective (Ex : site ministériel) ou subjective (Ex : page de blog).

Afin de disposer des éléments ci-dessus, nous avons exploité un dictionnaire de la langue française[77] converti en fichier Microsoft Access ainsi qu'un dictionnaire de valence de la langue française déjà décrit dans le chapitre consacré à l'indicateur de valence. Le dictionnaire utilisé comporte 289 576 entrées. Un extrait de ce dictionnaire est fourni *Tableau 32* :

[77] Ce disctionnaire est une production de l'Association des Bibliophiles Universels (ABU). Il est téléchargeable à l'adresse web http://abu.cnam.fr/DICO/mots-communs.html. Il est précisé dans la licence que « toute copie à des fins privées, à des fins d'illustration de l'enseignement ou de recherche scientifique est autorisée »

brut	lemmatise	type	Champ5
équipées	équipé	Adj	Fem+PL
équipées	équipée	Nom	Fem+PL
équipées	équiper	Ver	PPas+Fem+PL
équipement	équipement	Nom	Mas+SG
équipements	équipements	Nom	Mas+PL
équipent	équiper	Ver	IPre+PL+P3:SF
équipera	équiper	Ver	IFut+SG+P3
équiperaient	équiper	Ver	CPre+PL+P3

Tableau 32 : entrées du dictionnaire utilisé

Grâce à ce dictionnaire, nous pouvons donc identifier les adjectifs (le champ type comporte l'expression Adj), les verbes au conditionnel (la valeur de champ5 commence par CPre), les termes associés à l'expression d'une première personne du singulier (le champ brute contient je, moi, mon, mes, mien, miens, j'). Ce dictionnaire de départ a été traité à nouveau pour enlever les éléments ambigus. Si on considère le mot brut « équipées », ce terme est associé à la fois à un adjectif, un nom et un verbe. Seule une analyse du mot dans son contexte permettrait de connaître la désignation appropriée. A défaut d'une analyse plus fine, nous avons fait le choix de supprimer les éléments de la colonne brute qui donnaient lieu à plusieurs lignes dans le dictionnaire. En opérant ce retraitement nous sommes donc passés d'un dictionnaire de 289 576 entrées à un dictionnaire de 251 306 termes.

Le processus général de construction de l'indicateur de subjectivité est présenté *Figure 31*. Il comporte plusieurs étapes. On considère au départ un ensemble de pages web. Ces pages son aspirées grâce au logiciel WebPipe Pro. Le corpus ainsi obtenu est traité sous le logiciel TextPipe Pro. L'objectif est de recenser l'ensemble des termes de chaque page web ainsi que leur position (ordre) dans la page (*Tableau 33*).

ordre	page	mot
1	ecologie.caradisiac.com\France-a-quand-la-voiture-electrique-en-serie-139	france
2	ecologie.caradisiac.com\France-a-quand-la-voiture-electrique-en-serie-139	à
3	ecologie.caradisiac.com\France-a-quand-la-voiture-electrique-en-serie-139	quand
4	ecologie.caradisiac.com\France-a-quand-la-voiture-electrique-en-serie-139	la
5	ecologie.caradisiac.com\France-a-quand-la-voiture-electrique-en-serie-139	voiture
6	ecologie.caradisiac.com\France-a-quand-la-voiture-electrique-en-serie-139	électrique
7	ecologie.caradisiac.com\France-a-quand-la-voiture-electrique-en-serie-139	en
8	ecologie.caradisiac.com\France-a-quand-la-voiture-electrique-en-serie-139	série

Tableau 33 : exemple de fichier en sortie de TextPipe Pro

Ce prétraitement sous TextPipe Pro a pour effet de supprimer les éléments de ponctuation. Pour cette raison, nous n'avons pas pu compter le nombre de point d'exclamation par page. Nous nous privons donc d'une source d'information qui aurait pu être précieuse pour définir la nature subjective d'un texte.

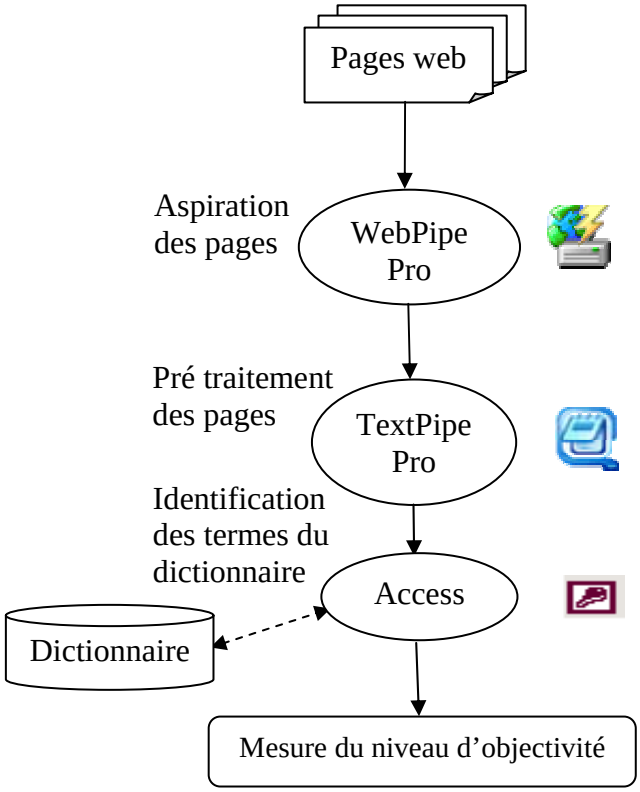


Figure 31 : Chaîne de traitement de l'information

Parmi les termes extraits par TextPipe Pro, nous identifions les caractéristiques de ceux qui sont présents dans le dictionnaire (adjectifs, verbes au conditionnel, tournure première personne…). Compte tenu du fait que les pages web n'ont pas toutes la même longueur, l'appréciation de ces éléments ne peut pas être effectuée dans l'absolu mais en relatif en rapportant chaque élément au nombre de termes de la page analysée rencontrés dans le dictionnaire.

3. Calibrage

Dans sa dimension opérationnelle, cette expérimentation qui a pour objectif de calibrer l'indicateur de subjectivité a été conduite en même temps que l'expérimentation qui avait pour objectif de valider l'indicateur de valence.

Pour cette raison, nous ne reprendrons pas ici le principe de détermination de l'échantillon des pages ainsi que la procédure de recueil de données qui nous a permis de récupérer les jugements des étudiants interrogés. En ce qui concerne l'indicateur de subjectivité, les étudiants étaient invités à exprimer une opinion sur une échelle à 4 modalités : très subjectif, plutôt subjectif, plutôt objectif, très objectif.

Le tableau fourni en annexe 7 comporte les jugements humains ainsi que les éléments qui vont nous permettre de calibrer l'indicateur de subjectivité. Le tableau présenté en annexe 8 est un sous-ensemble du tableau de l'annexe 7 qui reprend les 25 pages consensuelles sur lesquelles les internautes sont d'accord pour affecter un statut d'objectivité ou de subjectivité. Ces pages sont réparties en deux catégories : pages objectives et pages subjectives. On perd donc la granularité primitive de l'échelle à 4 modalités.
A partir de ce tableau, il est possible de calculer le taux d'incidence de chaque élément (je, conditionnel…) pour les pages objectives et subjectives afin d'identifier le ou les élément(s) discriminant(s). La synthèse des résultats est présentée *Tableau 34*.

	% adj	% mots au conditionnel	% je	% première personne du sing	% mots à valenceextremes	% valence<<0	% nous nos notre notres
pages estimées objectives par les juges	7,83	0,65	0,22	0,39	3,17	2,59	0,60
pages estimées subjectives par les juges	5,98	0,57	1,95	2,75	3,62	2,45	0,49

Tableau 34 : moyenne des valeurs par catégorie de pages

L'élément le plus discriminant pour juger du caractère objectif d'une page web est le nombre de formulations à la première personne du singulier dans la page considérée. Lorsque le texte de la page comporte plus de 1,5% de termes à la première personne du singulier (valeur de coupe), alors nous admettrons que cette page peut être considérée comme subjective ; dans le cas contraire, elle

sera considérée objective. Si on traduit ces résultats sous forme matricielle, on obtient le *Tableau 35*.

		Jugements humains		
		Page objective	Page subjective	total
Jugement automatique	Page objective	12	4	16
	Page subjective	0	9	9
	total	12	13	25

Tableau 35 : *Indicateur de subjectivité : indicateur automatique versus jugements humains perfectionnement*

Sur les 25 pages étudiées, il y a correspondance entre jugement automatique et humain dans 21 cas sur 25 soit 84% des cas. Pour les pages objectives, on observe une précision de 75% et un rappel de 100%.

Les valeurs théoriques qui seraient observées en cas d'indépendance des jugements humains et des valeurs calculées par l'indicateur sont présentées *Tableau 36*.

		Jugements humains		
		Page objective	Page subjective	total
Jugement automatique	Page objective	7,68	8,32	16
	Page subjective	4,32	4,68	9
	total	12	13	25

Tableau 36 : valeurs théoriques obtenues en cas d'indépendance des jugements humains

La somme des valeurs théoriques de concordance est de 12,36 (somme de la première diagonale) ce qui correspond à un taux de 0,4944 une fois rapporté au total. Ces deux valeurs permettent de calculer l'indicateur de Kappa qui est de 68,35%. L'accord est de bonne qualité au sens de Landis et Koch (1977).

Intuitivement, on peut pressentir que l'adresse url de la page web peut comporter des marqueurs qui pourraient permettre d'estimer avec une grande précision le caractère objectif ou subjectif d'une page web. Les pages dont le nom de domaine comporterait par exemple le terme « .gouv.fr » ont sans doute de grandes chances d'être qualifiées d'objectives. Il en va de même des pages du site Wikipedia. A l'inverse toutes les pages dont l'url contient le terme de blog ont sans doute plus de chance d'être qualifiées de subjectives par les évaluateurs. Nous avons pu vérifier ces éléments sur notre test de calibrage. Sur les 28 pages considérées, 4 sont issues de Wikipedia et 5 comportent le terme blog. On observe que les 4 pages issues de Wikipedia sont qualifiées d'objectives par les internautes. On constate aussi que 4 des 5 pages comportant le mot blog sont qualifiées de subjectives par les évaluateurs. L'intuition semble donc se vérifier sur ce petit exemple, même si le nombre de pages observées est trop réduit pour pouvoir en tirer une conclusion statistique. Il est difficile d'implémenter cet indicateur dans la mesure où la présence de ces éléments textuels dans l'url n'est pas systématique (ils apparaissent dans notre échantillon dans 9 cas sur 28). Cet indicateur pourrait intervenir à deux niveaux : en complément de l'indicateur principal défini en amont ou à titre confirmatoire.

On pourrait par exemple considérer que le statut défini sur la base de l'url prend l'ascendant sur l'indicateur principal. Cela signifierait par exemple que toute page web définie comme objective par l'indicateur automatique serait considérée comme subjective si elle comporte le terme blog dans l'url. Cela signifierait aussi que toute page web qualifiée de subjective par l'indicateur automatique serait considérée comme objective si elle comporte dans son adresse url les termes Wikipedia ou .gouv.fr. Les termes de blog, Wikipedia et .gouv.fr ne sont que des exemples qui ont pu être testés. Il est possible d'alimenter « en dur » une liste d'expressions caractéristiques figurant dans l'url de la page. On observe sur notre corpus que dans 8 cas sur 9, les pages qualifiées automatiquement d'objectives (ou subjectives) sont confirmées dans ce statut par l'indicateur basé sur l'url. Dans un cas sur 9 on observe une page définie comme objective par la machine et par les juges alors qu'elle comporte le terme de blog. Appliquer la règle précédente aurait donc conduit à dégrader l'indicateur de Kappa. L'observation du contenu de l'url et le fait de repérer

certaines expressions connues pour être objectives ou subjectives permet de renforcer la probabilité que la page soit effectivement objective ou subjective.

Les autres éléments comme le pourcentage d'adjectifs, le pourcentage de termes au conditionnel qui avaient été identifiés comme marqueurs potentiels de la siubjectivité d'un texte n'ont finalement pas été intégrés ni à titre individuel ni à titre de combinaison avec les autres éléments car ils dégradaient les indicateurs.

4. Test

L'objectif de cette seconde expérimentation est de tester l'algorithme construit dans l'étape précédente en aveugle sur un nouvel échantillon.

Cette seconde expérimentation a permis de tester l'indicateur de subjectivité ainsi que de calibrer l'indicateur de lisibilité (qui fera l'objet d'une présentation ultérieure). Nous avons considéré au départ un ensemble de 55 pages web (5 pages web pour un ensemble de 11 requêtes différentes). On demande à 100 étudiants de choisir deux requêtes et de noter les 5 pages web de chacune de ces deux requêtes. Pour chaque page qu'il qualifie, le juge doit choisir une modalité sur une échelle à 4 modalités pour la subjectivité.

Comme dans le test de calibrage, la dispersion des réponses fournies par les juges indépendants permet d'identifier des pages pour lesquelles les jugements sont consensuels. Le pourcentage de pages consensuelles obtenu est de 34 sur 55 soit 61%. Pour chaque page consensuelle retenue, nous avons limité les réponses des juges à deux modalités (subjective ou objective) et confronté ces jugements humains au jugement automatique.

L'indicateur automatique de subjectivité prend en compte le pourcentage de termes à la première personne du singulier dans le texte considéré. Nous avons opéré une double restriction portant, comme dans le test de calibrage, sur le nombre minimum de mots de la page web retrouvés dans le dictionnaire et un pourcentage minimum de mots à la première personne du singulier. Si on retient les pages comprenant plus de 400 mots dans le dictionnaire et si on affecte le statut de subjective aux pages dont 0,1% du texte est à la première personne du singulier, alors on obtient le *Tableau 37*.

		Jugements humains	
		Page objective	Page subjective
Jugement automatique	Page objective	16	3
	Page subjective	1	8

Tableau 37 : *Indicateur de subjectivité : indicateur automatique versus jugements humains perfectionnement 2*

Pour les pages objectives, on observe une précision de 84% et un rappel de 94%. Il y a concordance entre jugement humain et jugement automatique dans 24 cas sur 28 ce qui correspond à un pourcentage d'accord observé de 85%
Les valeurs théoriques qui seraient observées en cas d'indépendance des jugements humains et des valeurs calculées par l'indicateur sont présentées *Tableau 36*.

		Jugements humains		
		Page objective	Page subjective	total
Jugement automatique	Page objective	11,535714	7,4642857	19
	Page subjective	5,4642857	3,5357143	9
	total	17	11	28

Tableau 38 : valeurs théoriques obtenues en cas d'indépendance des jugements humains

La somme des valeurs théoriques de concordance est de 15,06 (somme de la première diagonale) ce qui correspond à un taux de 0,5378 une fois rapporté au total. Ces deux valeurs permettent de calculer l'indicateur de Kappa qui est de 69 %. L'accord est bon au sens de Landis et Koch (1977).

Il est intéressant de confronter l'appréciation des juges humains relative aux indicateurs de valence et de subjectivité. On peut observer Tableau 39 que sur les 23 pages non consensuelles au titre d'un des deux critères au moins, 2 seulement le sont au titre des deux critères. On observe aussi que les pages

négatives qui, au titre de l'indicateur de subjectivité, donnent lieu à une appréciation consensuelle sont considérées comme subjectives par les juges. Ces remarques réalisées sur un nombre limité d'observation mériteraient d'être étudiées dans un travail plus approfondi.

	non consensuelle	obj	sub	Total
non consensuelle	2	6	4	12
neg	4		5	9
neutre	3	4	3	10
pos	4	4	2	10
Total	13	14	14	63

Tableau 39 : ***Relation indicateur de valence indicateur de subjectivité***

5. Discussion des résultats et limites

Cette section a permis d'établir un état de l'art et de dégager des pistes d'indicateurs de subjectivité d'une page web. Ces pistes ont ensuite fait l'objet d'un test de calibrage et d'une expérimentation. L'indicateur qui a été retenu est le pourcentage « élevé » d'éléments à la première personne du singulier comme marqueur de subjectivité d'une page web. Cet indicateur a ensuite été confronté aux jugements humains dans une nouvelle expérience. Selon le test Kappa, l'indicateur de concordance est bon, signe que l'indicateur obtenu reflète bien les jugements humains. On peut adresser à cet indicateur un certain nombre de limites qui correspondent à des perfectionnements possibles.

- Le caractère très subjectif de la notion d'objectivité. La notion d'objectivité et de subjectivité est « fuyante ». Pour cette raison, nous avons choisi de raisonner sur un indicateur de subjectivité plutôt que d'objectivité. Pour valider les indicateurs que nous avons construits, nous avons confronté ces jugements automatiques à des jugements humains. Même si la subjectivité n'est pas comprise de la même façon pour tout le monde, on observe que les résultats des personnes interrogées vont dans le même sens bien souvent et qu'il est possible de définir des pages consensuelles. Les pages consensuelles sont des pages pour lesquelles l'écart type des réponses exprimées par les juges est faible. Cela correspond donc à une vision partagée de la subjectivité. Le pourcentage

de pages consensuelles est de 61% dans le test de subjectivité. Ces jugements humains ne caractérisent pas la subjectivité de la page mais la perception qu'ont les personnes interrogées de la subjectivité de la page.

- Limite liée à la nature de l'expérimentation conduite. D'autres pistes d'indicateurs avaient été évoquées mais elles n'ont pas résisté à l'épreuve des faits. Une d'entre elles consistait par exemple à considérer le nombre de termes à valence très négative dans la page considérée qui apparaît intuitivement comme corrélé à la subjectivité de la page web. Toutefois dans les tests qui ont été réalisés, ce critère n'est pas ressorti comme significatif. La raison est sans doute à chercher dans le fait que notre corpus est composé de plusieurs requêtes. Nous avons ainsi pu observer que les pages sur « l'énergie nucléaire » étaient souvent classées dans les pages subjectives car elles comportent, même si elles sont positives, une concentration forte de termes négatifs. Ces pages se retrouvaient donc classées dans les pages subjectives. Cet indicateur pourrait être utilisé pour comparer entre elles les pages d'un même univers (ce qui est le cas normal d'utilisation de la méthode). L'utilisation dans un domaine passe sans doute également par la construction d'un dictionnaire métier dans ce domaine.
- Nos indicateurs sont bien souvent définis sur une base textuelle. On a observé que la pertinence des indicateurs augmente avec le nombre de termes que l'on retrouve dans le dictionnaire pour la page web considérée. Ceci introduit donc une limitation de la méthode puisque toutes les pages du web ne satisfont pas à ces contraintes. L'indicateur lexical que nous avons utilisé permet surtout de qualifier des pages web informationnelles. Il est moins adapté à la caractérisation de pages web navigationnelles[78].
- Un biais méthodologique : dans ce travail, nous avons souhaité définir un indicateur de subjectivité d'un texte. Dans les expérimentations que nous avons conduites, la subjectivité est appréciée par la perception que peuvent en avoir des juges humains non professionnels. Or les juges humains peuvent baser leurs jugements de subjectivité sur des critères superficiels

[78] Les pages de navigation permettent à l'internaute d'affiner sa recherche et d'être redirigé vers des pages de contenu. Les pages de navigation permettent donc de passer de la page d'accueil du site à des niveaux de ramification plus fins.

comme par exemple le design du site web (Fogg, 2003). Il y a donc sans doute un écart entre le caractère subjectif d'un texte et la perception de cette subjectivité par les juges. On est alors face à un dilemme. Soit on souhaite mesurer le niveau de subjectivité d'une page web auquel cas il faut se doter d'une mesure de la subjectivité du texte et non pas de la perception de la subjectivité de ce texte. Soit on se résoud à proposer un indicateur de perception subjective. L'indicateur automatique retenu risque alors de cristalliser tous les biais, tous les éléments superficiels qui entrent en jeu pour définir la subjectivité d'un texte. L'indicateur retenu mesurerait alors davantage l'apparence de subjectivité d'une page web que sa subjectivité.

C. Niveau d'accessibilité

1. Etat de l'art de la notion d'accessibilité

a) Les accessibilités

L'accessibilité dans le domaine du web est définie par le W3C comme la capacité d'un site web à être utilisé par une personne ayant une incapacité quelconque. L'incapacité peut être de différentes natures :

- Elle peut être matérielle ou logicielle : ancienne version de navigateur, écran de petite taille, mauvaise connexion web.
- Elle peut être due à l'incapacité sensorielle ou cognitive de l'usager : déficit visuel, auditif, mobilité réduite, personne ne parlant pas ou parlant peu la langue.
- Elle peut être due à des contraintes environnementales : travail sous contraintes.

La notion d'accessibilité est donc loin d'être une notion qui signifie la même chose pour tout le monde. Un site web accessible pour un aveugle ne le sera pas forcément pour une personne évoluant sur un navigateur ancien. Cette notion d'accessibilité est donc contingente et difficile à satisfaire, certaines caractéristiques pouvant s'avérer difficilement compatibles entre elles.

b) Prise en compte institutionnelle

Cette problématique de l'accessibilité a été étudiée à la fin des années 1990 par le W3C qui a publié un guide d'accessibilité des contenus web[79]. Ce guide comporte de bonnes pratiques à respecter. En voici quelques unes :

- proposer une alternative textuelle à tout élément non textuel.
- ne pas faire reposer l'information sur la couleur.
- utiliser un langage naturel clair (penser aux outils de reconnaissance vocale, aux traducteurs).
- proposer un mécanisme de navigation clair.

Ce guide se décline autour d'un certains nombre de points de contrôle. Chaque point correspond à un niveau de priorité de type 1, 2, 3. Les éléments à niveau de priorité 1 correspondent à des éléments qui doivent être satisfaits au risque d'offrir un contenu en ligne inaccessible à plusieurs catégories de personnes. Les niveaux de priorité 2 et 3 correspondent à d'autres critères de moindre importance qui, s'ils sont satisfaits, lèvent des barrières à l'accessibilité.

A partir de ces éléments, le W3C a défini trois niveaux de conformité : A, double A et triple A.

- Le niveau A correspond à la satisfaction des éléments de niveau de priorité 1.
- Le niveau double A correspond à la satisfaction des niveaux de priorité 1 et 2.
- Le niveau triple A correspond à la satisfaction des niveaux de priorité 1, 2 et 3.

Les législateurs se sont également intéressés à cette question de l'accessibilité en la considérant avec un sens restreint d'accessibilité des ressources web aux handicapés.

Aux USA, le gouvernement a publié une loi (section 508) définissant les règles d'accessibilité d'un document web. Ces règles d'accessibilité des sites web aux handicapés doivent être respectées par les sites web fédéraux.

En France, le Comité Interministériel pour la Société de l'Information (CISI), précise le 10 juillet 2003 dans la mesure 2.6 le caractère obligatoire de

[79] http://www.w3.org/TR/WCAG10/#Conformance

l'accessibilité de l'information numérique publique aux handicapés dans le cadre de la révision de la loi de 1975 sur le handicap. L'article 47 de la Loi du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées précise que « Les services de communication publique en ligne des services de l'Etat, des collectivités territoriales et des établissements publics qui en dépendent doivent être accessibles aux personnes handicapées. » Dans la pratique, cette loi est rarement respectée.

c) Diagnostic d'accessibilité

Il existe un certain nombre d'initiatives qui ont pour objectif de faire évoluer des sites web vers plus d'accessibilité. En France, Accessiweb[80] a mis en place un système de labellisation payant. Pour obtenir le label, le site fait l'objet d'un audit où sont jugés 92 points regroupés en 13 catégories[81] :

1. Eléments graphiques
2. Cadres
3. Couleurs
4. Multimédia
5. Tableaux
6. Liens
7. Scripts
8. Eléments obligatoires
9. Structuration de l'information
10. Présentation de l'information
11. Formulaires
12. Aide à la navigation
13. Contenus accessibles

Il y a correspondance entre conformité aux règles d'accessibilité au sens du W3C et au sens d'Accessiweb : le label de bronze correspond au niveau de conformité A du W3C, le label d'argent au niveau de conformité double A et le label d'or au niveau de conformité triple A.

[80] http://accessiweb.org

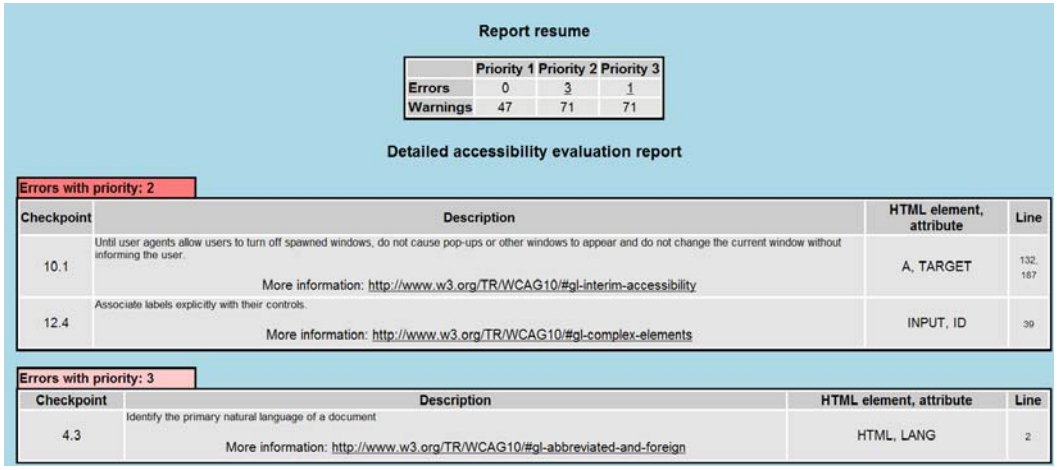
[81] Le détail de ces 92 critères est disponible http://www.accessiweb.org/_repository/files/id_26_1076666870250967.txt ou http://www.accessiweb.org/fr/Label%5FAccessibilite/criteres%5Faccessiweb/#fichiers

Il existe aussi un certain nombre d'outils logiciels dont l'objectif est de proposer un diagnostic d'accessibilité. L'accessibilité ne peut pas être uniquement testée par un logiciel : il existe dans les recommandations des éléments qui ne peuvent être examinés que par un diagnostic humain. Par exemple, le guide du W3C recommande que chaque image comporte une balise textuelle qui la définisse, conduisant à fournir une alternative textuelle à ceux qui n'ont pas accès aux images. Un outil va pouvoir s'assurer que la balise textuelle existe pour chaque image. Toutefois l'outil ne pourra pas s'assurer que le descriptif fourni dans la balise alt est bien en adéquation avec l'image. De la même manière, il peut être possible automatiquement de vérifier que le langage utilisé ne comporte pas trop d'abréviation ni de terme de jargon difficilement interprétables par certains traducteurs. Toutefois, un système automatique ne sera pas capable d'interpréter la clarté dans l'expression d'un texte. Cela fait de ces outils des éléments précieux mais qui doivent être complétés par l'expert humain.

Les outils de diagnostic, disponibles en ligne[82], offrent un premier niveau d'analyse. Chacun de ces outils se conforme aux recommandations du W3C et propose un diagnostic d'accessibilité. Eval Access 2.0[83] est l'un d'eux. Cet outil essaie de juger du niveau d'accessibilité au sens des niveaux d'accessibilité du W3C de type A, double A et triple A. Il propose un diagnostic détaillé et un tableau de synthèse. Le *Tableau 40* présente, à titre d'illustration, le diagnostic renvoyé à l'issue de la soumission à Eval Access 2.0 de la page d'accueil du site du laboratoire I3M (i3m.univ-tln.fr). Ce diagnostic a été effectué le 18 février 2008.

[82] http://www.w3.org/WAI/ER/tools/complete

[83] disponible à l'adresse http://sipt07.si.ehu.es/evalaccess2/index.html



Report resume

	Priority 1	Priority 2	Priority 3
Errors	0	3	1
Warnings	47	71	71

Detailed accessibility evaluation report

Errors with priority: 2

Checkpoint	Description	HTML element, attribute	Line
10.1	Until user agents allow users to turn off spawned windows, do not cause pop-ups or other windows to appear and do not change the current window without informing the user. More information: http://www.w3.org/TR/WCAG10/#gl-interim-accessibility	A, TARGET	132, 167
12.4	Associate labels explicitly with their controls. More information: http://www.w3.org/TR/WCAG10/#gl-complex-elements	INPUT, ID	39

Errors with priority: 3

Checkpoint	Description	HTML element, attribute	Line
4.3	Identify the primary natural language of a document More information: http://www.w3.org/TR/WCAG10/#gl-abbreviated-and-foreign	HTML, LANG	2

Tableau 40 : résultat de la soumission à Eval Access de la page d'accueil du site i3m.univ-tln.fr

Le *Tableau 40* permet de visualiser de manière synthétique (dans le tableau du haut) le nombre d'erreurs et d'alertes pour chaque niveau de conformité. Les tableaux du dessous permettent de reconstituer, pour chaque niveau de conformité, le nombre de fois où chaque erreur différente a été observée dans la page web considérée.

Il existe enfin certains cabinets de conseil spécialisés sur cette question. Le cabinet Temesis – région Aquitaine[84], par exemple, a mis en ligne une liste de bonnes pratiques en matière de qualité de site web et propose une méthodologie itérative permettant à un site de progresser sur le chemin d'une meilleure accessibilité.

2. L'accessibilité comme facette : protocole et mise en œuvre calculatoire

Proposer une facette accessibilité suppose d'être capable d'affecter à une page web une valeur quantitative permettant de qualifier son niveau d'accessibilité. Cette valeur doit être calculée automatiquement à partir d'éléments mesurés en ligne par un logiciel. Cette mesure automatique de l'accessibilité bien qu'incomplète permet d'éliminer certaines pages web qui ne sont pas accessibles puisqu'elles ne satisfont pas aux critères automatiques. A contrario,

[84] http://www.temesis.com/ Elie Sloim

l'obtention d'un bon score automatique ne permet pas de préjuger de l'accessibilité ou non d'une page web, beaucoup d'autres points étant à vérifier humainement. Nous formulons à ce stade deux hypothèses :

- H1 : si un site satisfait aux critères automatiques, il est probable qu'il satisfasse également aux critères d'évaluation humains.
- H2 : si une page du site a un bon score d'accessibilité, il est probable que le site obtienne lui aussi un bon score.

Nous n'avons pas trouvé d'outil qui fournisse un indicateur quantitatif d'accessibilité. Les outils qui existent s'inscrivent davantage dans une démarche de diagnostic de l'accessibilité permettant au site de s'améliorer plus que dans l'évaluation sommative par un indicateur unique.

Notre objectif va maintenant consister à décrire l'indicateur d'accessibilité que nous avons construit. Pour construire cet indicateur, nous exécutons Eval Access 2.0 pour un corpus de pages web. Le logiciel nous renvoie alors trois familles d'informations :

- le niveau de priorité : A, double A, triple A
- le nombre d'erreurs différentes pour un niveau de priorité donné
- le nombre de fois où une erreur a été commise.

Nous nous basons sur ces indices pour construire l'indicateur d'accessibilité.

a) Construction d'un indicateur d'accessibilité

Il n'y a pas de méthode universelle pour calculer une note à partir de divers indicateurs, toutefois, nous proposons les règles de construction suivantes :

- Chaque niveau de priorité (A, AA, AAA) est évalué sur 30 points.
- A chaque nouvelle infraction, 5 points sont retirés à la note du niveau.
- A chaque infraction déjà mentionnée, 1 point est retiré à la note du niveau.
- Si la décote excède le quota du niveau, la note de 0 est attribuée au niveau.
- Une note globale sur 20 est calculée par pondération des notes des trois niveaux de sorte que l'évaluation de A soit sur 12, AA sur 5 et AAA sur 3.

Formule générale de calcul de la note totale d'accessibilité :

$$Note_{totale} = \frac{12Note_A + 5Note_{AA} + 3Note_{AAA}}{30}$$ en notant pour A :

- $\mathbf{n}_A$ le nombre de **n**ouvelles infractions
- $\mathbf{m}_A$ le nombre d'infractions déjà **m**entionnées
- $Note_A = Max\{30 - m_A - 5n_A ; 0\}$ la note du niveau A

b) Validation de l'indicateur d'accessibilité

Nous avons validé cet indicateur à partir d'un corpus composé de pages validées Accessiweb et d'un corpus d'autres pages correspondant aux pages d'accueil des entreprises du CAC40. Les résultats sont fournis Annexe 15.
La moyenne de l'indicateur trouvé pour les entreprises Accessiweb est de 16.46/20. La moyenne des entreprises du CAC 40 est de 10.76/20.

Remarques :

- Cet indicateur ne prend pas en compte la taille de la page (une grande page est susceptible de comprendre davantage d'infractions).
- Il a tendance à valoriser les pages simples sans artifices.
- Les pondérations choisies permettent de discriminer les pages que nous avons consultées. Toutefois, d'autres pondérations pourraient être utilisées.

3. Test

Il n'est pas envisageable de tester la pertinence de l'algorithme en confrontant la mesure automatique de l'accessibilité à une mesure humaine. Nous avons testé le modèle à travers une comparaison entre le score moyen obtenu par un groupe de pages web disposant d'un label d'accessibilité et un groupe de page web ne disposant pas de ce label.

Nous avons ainsi appliqué notre indicateur à un ensemble de 21 pages brésiliennes prises aléatoirement dans un ensemble de 385 pages constituant, en février 2008, le répertoire national des sites accessibles au Brésil. Ce répertoire est accessible en ligne[85]. La moyenne de notre indicateur d'accessibilité sur ces pages est de 16.44/20. Les résultats sont fournis annexe 14. En comparaison, le score obtenu par les pages d'accueil des sites des entreprises du Cac 40 français (annexe 15) est de 10.76/20.

[85] http://www.dasilva.org.br/

4. Discussion des résultats

L'estimateur d'accessibilité que nous avons construit nous a permis de discriminer les sites labellisés des sites non labellisés. Or il est possible que certains sites non labellisés possèdent les qualités d'un site accessible. Pour cette raison notre mesure minore sans doute les écarts qui auraient pu être observés entre des sites accessibles et des sites qui ne le sont pas.

Comme mentionné précédemment, d'autres pondérations auraient pu être utilisées donnant par exemple plus de poids à tel niveau de priorité ou alors à telle catégorie d'infraction. Toutefois, il est important de garder en tête que la note relative de deux pages prime sur la note absolue. Le choix pourra être fait ensuite de la coupe : à partir de quelle note on considère que la page est accessible.

Il pourrait être souhaitable, en effet, de déterminer dans une nouvelle étude le seuil (la coupe) au-delà duquel on considère qu'un site est accessible. Pour ce faire, on pourrait utiliser la même démarche que celle expliquée dans ce travail pour la polarité en utilisant le coefficient de Kappa (pré calibrage suivi de test).

Il serait également souhaitable d'estimer plus finement notre indicateur (qui est quantitatif et non binaire à l'origine) en demandant par exemple à un expert (ou plusieurs) de classer divers sites du plus au moins accessible. Nous pourrions alors comparer les rangs obtenus à ceux proposés par l'estimateur automatique (en utilisant le test de corrélation non paramétrique de Spearman).

L'indicateur proposé a le mérite d'être simple à comprendre et facile à implémenter (en utilisant les résultats fournis par Eval Access 2.0).

Nous aurions pu aussi retenir une dimension plus étroite de l'accessibilité en nous intéressant par exemple à l'aspect visuel proposé à l'internaute. Le W3C a formulé des recommandations relatives au contraste entre couleur d'arrière

plan et couleur de premier plan. Accesskeys[86] a implémenté ces préconisations. Pour se conformer aux préconisations du guide d'accessibilité web, le premier plan et l'arrière plan doivent produire un contraste suffisant pour être vu par une personne ayant un déficit visuel.

Pour vérifier les conditions d'accessibilité les définitions suivantes sont utiles ;

- Définition suggérée par le W3C pour calculer la luminosité d'une couleur RGB (R,G,B) : luminosité = (299 R + 587 G + 114 B) / 1000
- Définition suggérée par le W3C de la différence entre deux couleurs : RGB (R_1,$G_{1,}$,B_1) et RGB (R_2,$G_{2,}$,B_2) : diff = $|R_2-R_1| + |G_2-G_1| + |B_2-B_1|$

L'expérience a montré que :

- Deux couleurs ont une bonne visibilité si leur différence de luminosité est supérieure à 125.
- La différence entre la couleur de fond et celle de premier plan doit être supérieure ou égale à 500 pour que la page soit considérée accessible.

Cet indicateur n'a pas été retenu. La dimension visuelle est intéressante mais elle est déjà intégrée dans l'indicateur que nous avons observé.

D. Niveau de lisibilité

1. Etat de l'art de la notion de lisibilité :

Selon Henry (1975), la lisibilité désigne "le degré de difficulté éprouvé par un lecteur essayant de comprendre un texte". La lisibilité peut être définie comme "une aptitude du texte à se faire comprendre" (Bourque, 1989). Pour Gélinas et al. (1993), la lisibilité est l'aptitude d'un texte à être lu rapidement, compris aisément et bien mémorisé. Pour Gélinas et al. (1993), la notion de lisibilité est différente de celle d'intelligibilité. La notion de lisibilité renvoie aux caractéristiques micro-structurelles d'un texte. La lisibilité peut être mesurée à l'aide d'indices de surface du texte que sont par exemple les mots et les phrases. Le concept de lisibilité ne concerne donc pas la structure argumentative d'un texte. L'intelligibilité s'intéresse davantage au lien entre les idées.

[86] disponible à l'adresse url : http://www.accesskeys.org/tools/color-contrast.html

Dans la littérature, les indicateurs de lisibilité sont utilisés dans le domaine de la pédagogie et également pour permettre aux pouvoir publics de produire des contenus d'information qui puissent être compris par tous (Gélinat et al 1991). Le développement d'indicateurs de lisibilité présente également un intérêt pour permettre l'accès à des contenus francophones à des personnes pour lesquelles le français n'est pas la langue native.

On peut distinguer avec Mesnager trois éléments qui permettent de définir la lisibilité d'un texte.

- la typographie : Richaudeau (93) a montré que pour des textes disponibles sur support papier, l'espacement, l'interlignage jouent un rôle important. Le découpage du document en blocs favorise la compréhension de la structure du texte. De ce point de vue, la lisibilité des textes multimédias obéit à des règles particulières.
- Le lexique : On considère que plus un texte comporte de mots rares, plus son contenu sera réputé difficile.
- la syntaxe de la phrase : On considère que plus une phrase est longue, plus les mots qu'elle comporte ont de lettres ou de syllabes, plus la phrase est complexe.

Dans la littérature, la lisibilité d'un texte est appréhendée par différents indicateurs. Ces indicateurs accordent plus ou moins d'importance au facteur lexical ou syntaxique. Voici quelques-uns de ces indicateurs parmi les plus employés.

- Formule de Kincaid : 11.8 * syllabes/mot + 0.39 * mots/phrases - 15.59
  Cette formule a été développée par l'armée américaine pour classer les supports de cours de l'armée par difficulté de lecture selon une échelle de 5,5 à 16,3.
- Automated Readability Index : 4.71 * caractères/mots + 0.5 * mots/ phrase - 21.43
- Formule de Coleman-Liau (1975) : Lisibilité = 5.89 * nombre de caractères par mot - 0.3 * nbre Phrases / (100*nbre mots) - 15.8

- Flesch Index = 206.835 - 84.6 * syllabes/mots - 1.015 * mots/phrases
  Cet indicateur a été développé par Flesh (1948). Il a été construit à partir de textes scolaires. Cet indicateur est compris entre 0 (difficile) et 100 (facile).
- Kandel et A. Moles (1958) ont proposé une adaptation française de l'indicateur de Flesh : Facilité de lecture = 209 - (0,68 x W) - (1,15 x S). où W désigne le nombre moyen de syllabes pour 100 mots et S le nombre moyen de mots par phrase.

Le *Tableau 41* définit les niveaux de lecture en fonction de l'indice de Kandel et Moles.

Indice de 0 à 100	Niveau de difficulté	Genre de support	% de la population ayant accès
0-30	Très difficile	Scientifique	4,50%
30-50	Difficile	Technique	33%
50-60	Assez difficile	Vulgarisation	54%
60-70	Normal	Rewriting de masse	83%
70-80	Assez facile	Revues familiales	88%
80-90	Facile	Poches détective	92%
90-100	Très facile	Bandes dessinées	93%

Tableau 41 : Niveau de difficulté correspondant à l'indice de Kandel et Moles

- Selon Gunning (1952), la lisibilité repose essentiellement sur la longueur des phrases et des mots. Sa formule retient deux facteurs : le nombre moyen de mots par phrase (L) et la fraction de mots de plus de trois syllabes (M) : Fog Index = (L + 100M) x 0,4
- La formule de Dale et Chall (1948) : $I = 0,15\ x_1 + 0,04\ x_2$ où x_1 représente le facteur lexical défini comme le pourcentage des mots du texte absents d'un vocabulaire de base et x_2 correspond au nombre moyen de mots par phrase.
- Henry (1975) a proposé 3 formules de lisibilité adaptées au français. La formule la plus simple dite « manuelle » retient trois variables : le

nombre de mots par phrase, le pourcentage de mots différents d'une liste de mots usuels (la liste de Gougenheim : 3500 mots de base de la langue française) et le pourcentage de mots dits « marques de dialogue ».

Le *Tableau 42* récapitule les indicateurs et l'importance qu'ils accordent respectivement aux dimensions lexicales et syntaxiques.

	Dimension lexicale	Dimension syntaxique
Formule de Kincaid		X
Automated Readability Index		X
Coleman-Liau		X
Flesch Index		X
Kandel et A. Moles (adapté à la langue française)		X
Gunning : Fog Index		X
Dale et Chall	X	X
Georges Henry (adapté à la langue française)	X	X

Tableau 42 : Synthèse des indicateurs de lisibilité

B. Labasse (1999) écrit au sujet de ces indicateurs : « On pourrait trivialement comparer un indicateur de ce type au voyant d'un tableau de bord, qui peut alerter sur un incident, mais non prouver, en restant éteint, que le véhicule ne connaît pas de dysfonctionnement d'une autre nature... »

Le Tableau 43 permet d'identifier des outils en ligne qui permettent de calculer certains de ces indicateurs

	* Readability index calculator	** Readability.info	***Textalyser
Formule de Kincaid	X	X	
Automated Readability Index		X	
Coleman-Liau		X	
Flesch Index	X	X	
Kandel et A. Moles (adapté à la langue française)	X		
Gunning : Fog index		X	X
Dale et Chall			
Georges Henry (adapté à la langue française)			

Tableau 43 : outils en ligne disponibles pour calculer les indicateurs de lisibilité

*http://www.standards-schmandards.com/exhibits/rix/index.php

** http://www.readability.info/info.shtml

*** http://textalyser.net/index.php?lang=en#analysis

Les outils disponibles en ligne peuvent être testés à partir de textes copiés collés ou d'une url saisie.

Redish et Seizer (1985) ont exprimé de grandes réserves quand à l'utilisation de ces formules de lisibilité. Plusieurs raisons nous conduisent à la prudence quand au recours à ces indicateurs.

- Ces indicateurs ont été conçus pour la langue anglaise. Leur transposition directe au français n'est pas toujours envisagée. De ce point de vue, seul l'outil « Readability index calculator » intègre un indicateur calibré pour les pages francophones.
- Les outils que nous avons pu tester librement sur le web s'en tiennent à une définition très limitée de la lisibilité et apprécient la lisibilité à travers des éléments de type syntaxique : taille des mots, taille des phrases, nombre de mots par phrase. La dimension lexicale n'est jamais prise en compte.

- Les outils reposant sur une analyse lexicale posent problème car la communication sur le web échappe parfois aux règles syntaxiques d'usage. On rencontrera par exemple peu de « . » sur un page web comportant de nombreux liens vers une autre page. Si le « . » est utilisé pour compter le nombre de phrases et en déduire le nombre de mots par phrase, l'ordinateur risque de renvoyer un indicateur erroné.

2. La lisibilité comme facette : protocole et mise en œuvre calculatoire

Nous proposons de considérer au départ deux indicateurs de lisibilité représentant les deux grandes familles de lisibilité au sens lexical et syntaxique. Ces deux indicateurs seront confrontés au jugement d'un groupe d'internautes.

La lisibilité syntaxique sera appréciée à travers la formule de Kandel et Moles. Cette formule est mise en œuvre par l'outil « Readability index calculator ». Le travail consiste alors à soumettre en boucle à l'interface web de « Readability index calculator » les url des pages de notre corpus.

La lisibilité lexicale vise à appréhender la richesse du vocabulaire du texte. Le calcul de cet indicateur passe par la mise en œuvre d'une chaîne de traitement semi automatique de l'information. La première étape consiste à décomposer la page web à analyser en ses différents mots. Nous utilisons pour cela l'outil TextPipe Pro. En entrée, cet outil accepte une adresse url. En sortie il fournit un fichier texte sur chaque ligne duquel figure de manière ordonnée un mot de ce texte. Le fichier ainsi obtenu est soumis à un dictionnaire qui comporte pour chaque mot de la langue française sa fréquence d'apparition dans notre langue. Nous avons utilisé le dictionnaire de l'Université Descartes à Paris[87] . Ce dictionnaire mentionne la fréquence de chaque mot français. Dans ce dictionnaire, deux estimateurs des fréquences d'usage des termes du vocabulaire français sont proposés. Le premier est fondé sur un sous-ensemble de romans tirés du corpus FRANTEXT. Le second repose sur un corpus de sous-titres de films très récents ce qui permet d'avoir du vocabulaire actuel. Le

[87] http://www.lexique.org/telecharger.php

corpus de films fait ressortir des éléments du langage parlé tels que Salut, Bonjour, Au revoir, Oui, ou Non. Chaque élément de vocabulaire est défini par 4 indicateurs :

1_ortho	2_phono	3_lemme	4_cgram	5_genre	6_nombre	7_freqlemfilms	8_freqlemlivres	9_freqfilms	10_freqlivres
dansant	d@s@	danser	VER			108.14	92.57	2.34	5.54
dansante	d@s@t	dansant	ADJ	f	s	1.65	6.89	0.48	1.76
dansantes	d@s@t	dansant	ADJ	f	p	1.65	6.89	0.21	1.96
dansants	d@s@	dansant	ADJ	m	p	1.65	6.89	0.37	0.61
danse	d@s	danse	NOM	f	s	41.06	35.14	38.62	29.19
danse	d@s	danser	VER			108.14	92.57	18.46	9.8
dansé	d@se	danser	VER	m	s	108.14	92.57	5.27	4.32
dansée	d@se	danser	VER	f	s	108.14	92.57	0.11	0.27
dansent	d@s	danser	VER			108.14	92.57	3.14	5.54

Freqlemfilms : la fréquence du lemme selon le corpus de sous-titres de films (par million d'occurrences), sachant qu'il comprend 16.6 millions de mots ;

Freqfilms : la fréquence du mot selon le corpus de sous-titres de films (par million d'occurrences) ;

Freqlemlivres : la fréquence du mot selon le corpus de livres (par million d'occurrences).

Freqlivres : la fréquence du lemme selon le corpus de livres (par million d'occurrences) sachant qu'il présente 14.7 millions de mots.

Pour évaluer des corpus web, il aurait été plus judicieux de se constituer une base de vocabulaire type du web (à partir d'un échantillon de pages représentatives du web). En effet, certains termes risquent d'être fortement surreprésentés dans les pages web par rapport à des textes de films ou de livres.

Voici les hypothèses qui ont été introduites pour construire cet indicateur :

H1 : l'indicateur est construit à partir de la fréquence du terme dans le corpus de film pour les raisons sus évoquées.

H2 : l'indicateur est construit à partir des valeurs non lemmatisées. En effet une même expression lemmatisée (verbe être) peut donner lieu à des formes non lemmatisées plus ou moins occurrentes (sommes et fussiez).

H3 : l'occurrence d'un terme dans le corpus de films est une indication de son niveau d'accès. Plus un terme est d'usage courant, plus la page web qui le contiendra sera considérée comme lisible.

H4 : une page lisible comportera des mots lisibles.

3. Calibrage

Dans un premier temps, nous présentons la méthode de calibrage en utilisant le même jeu de données que celui utilisé pour l'indicateur de subjectivité. Ce jeu comporte 11 requêtes avec pour chacune d'elle 4 modalités. Le juge est un étudiant qui doit noter 10 pages web correspondant à deux requêtes qu'il choisit. Le juge doit préciser s'il trouve la page très facile, plutôt facile, plutôt difficile ou très difficile. Un premier filtre permet d'extraire les pages pour lesquelles il n'y a pas de consensus entre les juges. On procède de la même manière que précédemment en supprimant les pages pour lesquelles les modalités de réponses exprimées par les juges ont un écart type trop élevé. A l'issue de ce processus, il reste 44 pages consensuelles sur 55 soit 80 %. Il est à noter que pour le même jeu de données avec les mêmes juges, l'indicateur de subjectivité, également évalué sur 4 modalités, avait donné lieu à un pourcentage de pages consensuelles de 61%. Les juges semblent avoir plus de facilité à se mettre d'accord sur un critère de lisibilité que sur un critère de subjectivité.

Nous avons successivement soumis ce corpus de pages à deux indicateurs différents. Le premier s'appuie sur la formule de Kandel et Moles. Le second prend en compte une vision de la lisibilité qui dépend de la richesse du vocabulaire utilisé.

L'indicateur de Kandel et Moles a été calculé sur chacune des pages consensuelles. Pour rendre les résultats cohérents avec ceux de l'évaluation humaine, les valeurs de l'indicateur automatique sont regroupées autour de trois modalités nommées FAC (pour Facile) MOY (pour moyen) et DIFF (pour Difficile). Une page est estimée facile si son indicateur de Kandel et Moles est supérieur à 70. Une page est jugée de difficulté moyenne si elle a un indicateur compris entre 50 et 70. Une page dont l'indicateur est inférieur à 50 est considérée comme difficile. Le *Tableau 44* visualise la comparaison entre le statut des pages donné par les juges et le statut affecté grâce à l'indicateur de Kandel et Moles.

Lisibilité définie par les juges

		Difficile	Moyen	facile
Lisibilité Kandel et Moles	Difficile	2	7	2
	Moyen	2	19	6
	Facile	0	4	1

Tableau 44 Comparaison entre le statut de lisibilité défini par les juges et par l'indicateur de Kandel et Moles

Pour les pages difficiles, on observe une précision de 18% et un rappel de 50%. Le pourcentage d'accord observé est de 51%. En comparant le *Tableau 44* au tableau des valeurs théoriques qui prend en compte l'indépendance des jugements, on obtient un indicateur de Kappa de 0,049. La concordance entre les valeurs exprimées par les juges et les valeurs calculées par l'indicateur de Kandel et Moles est donc mauvaise.

Une des explications de ce mauvais résultat se trouve sans doute dans le fait que l'indicateur retenu est peu adapté à des données web. Le nombre de phrases (qui est une des variables de cet indicateur) est calculé à partir d'un comptage des marques de ponctuation. Dans les pages web de navigation vers d'autres pages, on trouvera comparativement peu de marques de ponctuation ce qui va conduire à augmenter artificiellement la valeur de l'indicateur et donc à rendre les pages plus difficiles que ce qu'elles sont réellement.

Compte tenu du fait que le nombre de phrases est difficile à apprécier sur le web, nous avons choisi de ne retenir dans la formule de Kandel et Moles que la première partie qui considère le nombre moyen de syllabes par mot. Nous n'avons pas calculé le nombre de syllabes de chaque mot mais uniquement des mots non vides. Une liste de 423 mots vides (ou mots communs) a été obtenue en considérant les termes du dictionnaire de l'Université Paris Descartes pour lesquels la fréquence moyenne d'apparition dans le corpus de film est supérieure à 100. Cette liste de mots vides est donnée en annexe 12. On a considéré que les pages web qui avaient un nombre moyen de syllabes par mot inférieur à 2,245 étaient faciles, que les pages web qui avaient un nombre moyen de syllabes par mot supérieur à 2,79 étaient difficiles et que les autres étaient de difficulté moyenne. On peut ainsi comparer le statut des pages calculé par l'indicateur et le statut affecté par les juges. Les données sont présentées *Tableau 47*.

		Lisibilité définie par les juges		
		Difficile	Moyen	facile
Lisibilité Kandel et Moles	Difficile	1	4	1
	Moyen	3	24	3
	Facile	0	4	5

Tableau 45 Comparaison entre le statut de lisibilité défini par les juges et par le nombre moyen de syllabes par mot

Pour les pages difficiles, on observe une précision de 16% et un rappel de 25%. Le pourcentage d'accord observé est de 30 sur 45 soit 66%.
La valeur de Kappa obtenue est 0,29 ce qui correspond à un indicateur médiocre. L'indicateur calculé grâce au nombre de syllabes est bien plus satisfaisant que l'indicateur de Kandel et Moles mais il reste trop faible pour être acceptable.

Cet indicateur de lisibilité introduit un autre biais lié à une forme d'effet de démonstration du lecteur : lorsqu'un juge considère qu'une page est subjective ou objective, il ne se juge pas lui-même. Lorsque ce même juge attribue à la page le statut de page difficile, il reconnaît qu'il a du mal à en comprendre le sens. Pour des raisons d'organisation, les étudiants qui ont participé à ce test de calibrage étaient dans une petite salle dans des conditions de promiscuité. Il est donc possible que les étudiants aient déclaré des pages plutôt faciles alors qu'elles ne l'étaient pas de peur d'être mal considérés par leurs camarades. Il conviendra dans le test réel de procéder à une évaluation en respectant l'anonymat du juge.

Nous avons soumis le même jeu de données à un autre indicateur qui s'appuie sur la fréquence d'apparition du vocabulaire de la page web dans un corpus de films. Nous ne disposons pas d'indicateur de référence. Les données du dictionnaire disponible présentent une grande dispersion des fréquences. Il y a un rapport de plus de 400 000 entre le terme le plus fréquent dans le corpus (je avec une fréquence de 26198) et le terme le moins fréquent (maroquinerie dont la fréquence est de 0,06). Cette amplitude nous a conduits à supprimer de

l'analyse les 423 mots vides dont la fréquence d'apparition est supérieure à 100. On a calculé la fréquence moyenne d'apparition des mots de la page web à qualifier dans ce dictionnaire. On considère que lorsque la moyenne des fréquences de la page est inférieure à 12, la page est difficile. Elle est facile lorsque la fréquence moyenne est supérieure à 18,5 et de difficulté moyenne entre ces deux limites.

Ces valeurs nous permettent de catégoriser cette variable et de procéder au test de concordance en comparant le jugement de lisibilité humain et le jugement automatique. Les résultats sont fournis *Tableau 46*.

		Lisibilité définie par les juges		
		Difficile	Moyen	facile
Lisibilité Kandel et Moles	Difficile	0	2	2
	Moyen	3	23	6
	Facile	1	7	1

Tableau 46 Comparaison entre le statut de lisibilité défini par les juges et par la fréquence d'apparition moyenne des mots de la page

Le pourcentage d'accord observé est de 24 sur 45 soit 53%. L'indicateur de Kappa vaut –0,04 ce qui correspond à une très mauvaise concordance. Ce résultat est peut être en partie dû au fait que le test de calibrage a porté non sur des pages d'un même thème mais sur des pages appartenant à 11 thèmes différents. Il faudrait donc pouvoir distinguer la lisibilité inter requête de la lisibilité intra requête. Nous avons demandé aux juges de noter la lisibilité de 5 pages d'une même requête. Ce travail leur a permis de noter les 5 pages les unes par rapport aux autres. Il semble difficile de comparer la lisibilité de pages ne correspondant pas à la même requête. Cela signifie qu'il faudrait, dans le test ultérieur, pouvoir considérer un ensemble homogène de pages sur un thème donné.

4. Conclusion et perspectives

A l'issue de ce paragraphe, nous n'avons pas pu identifier d'indicateur de lisibilité offrant une concordance satisfaisante. L'indicateur le plus efficace de ceux qui ont été testé est celui qui repose sur le nombre moyen de syllabes

d'une page web. Cet indicateur arrive à une prédictibilité de la valeur des juges dans 66% des cas. Il y a plusieurs pistes de développement d'un indicateur de lisibilité : la plus pertinente consiste à raisonner sur la fréquence d'utilisation des termes dans les pages web et non dans un corpus de film.

L'expérimentation réalisée dans le test de calibrage révèle le biais associé à une forme d'effet de démonstration. Il n'est en effet pas sur que la déclaration de la lisibilité faite par le juge reflète la lisibilité qu'il a percue ni la lisibilité réelle du document. En effet, le juge a sans doute peur d'être jugé négativement s'il considère qu'un document est peu lisible. Pour gommer cet effet de démontsration, au moins partiellement, il serait possible d'interroger les juges individuellement et de manière anonyme. Sans doute y-a-t-il d'autres biais à l'œuvre. Pour ces raisons, nous avons choisi de ne pas effectuer de test car nous ne pensons pas que la pertinence de l'indicateur de lisibilité puisse s'effectuer à l'aune d'une comparaison avec la perception de lisibilité déclarée par les internautes.

E. Niveau de fraîcheur

Il y a une légitimité, dans une logique d'Intelligence Economique à s'intéresser à un indicateur de fraîcheur ou de régularité avec laquelle la page web est mise à jour. Cela permet d'identifier des sources d'information vivantes et à l'opposé, celles qui sont mortes. Fetterly et al. (2003) ont consacré une large étude à la présentation du suivi de l'évolution de pages web. Ntoulas et al (2004) ont montré que parmi les pages web qui subsistent d'une année sur l'autre, la moitié d'entre elles n'ont pas changé. Si on observe les changements observés sur une semaine, 70% des changements correspondent à moins de 5% de modifications par rapport à la version initiale. Tous ces résultats sont dépendants de la maturité du web et ces chiffres ont sans doute évolués depuis 4 ans.

L'indicateur de fraîcheur suppose le suivi régulier d'un corpus de pages web et de la mesure de l'écart entre chacune des pages entre deux captures. Cette facette est intéressante car elle correspond à un besoin qui est mal couvert par les outils de recherche généralistes actuels : les moteurs de recherche permettent tous de faire une recherche dans des pages qui ont été modifiées

depuis 1 semaine, 1 mois… mais aucun moteur ne va privilégier la régularité de mise à jour de la source d'information.

Il est difficile d'envisager de déporter le calcul de cet indicateur sur un outil existant. Le moteur de recherche « web.archive.org » retrace les différentes modifications d'un site dans le passé mais ce moteur ne recouvre pas toutes les pages du web. De plus, ce moteur ne précise pas les changements réalisés durant la dernière année. Enfin, cet outil ne précise pas le taux de modification de la page web (est-ce une modification mineure ou une modification de fond).

Un certain nombre d'outils d'alerte permettent d'informer lorsqu'une source d'information web a été modifiée. Toutefois, ces alertes fonctionnent souvent par envoi de mail régulier ce qui implique une complexité de récupération des données lorsqu'on envisage le suivi d'un corpus de pages web important. De plus, ces outils d'alerte disposent rarement d'indicateurs permettant de chiffrer le taux de modification de la page. Or, c'est ce genre d'indicateur de synthèse qui nous intéresse.

Cet indicateur n'a pas été mis en œuvre pour des raisons calculatoires. Le mettre en œuvre exigerait des ressources importantes et un recul du temps pour être capable de suivre en dynamique l'évolution d'un corpus de pages web.

F. Facette construite à partir d'informations relationnelles

Il peut sembler curieux d'introduire une facette correspondant à la dimension relationnelle qui est déjà privilégiée par les indicateurs de pertinence des moteurs de recherche. Le risque est de développer un indicateur relationnel redondant de celui du moteur de recherche.

Dans ce travail, la prise en compte de la dimension relationnelle répond à trois motivations :

- Aujourd'hui, l'indicateur de pertinence des moteurs de recherche est opaque comme cela a été souligné précédemment. Nous souhaitons privilégier un indicateur de centralité décrit dans la littérature sur le sujet (centralité d'intermédiarité par exemple).

- Alors que l'indicateur relationnel de type PageRank de *Google* est construit sur l'ensemble des pages web[88], notre indicateur est construit sur un espace web contextualisé correspondant aux réponses à une requête.
- L'analyse relationnelle ne doit pas être écartée d'un nouvel indicateur de pertinence dans lequel l'internaute aura le loisir de mobiliser ou pas la facette relationnelle.

La démarche que nous proposons part de l'analyse des interactions hypertextuelles entre un corpus de pages web renvoyées suite à l'interrogation d'un moteur de recherche. L'objectif est de produire des indicateurs de niveau de centralité. Ces indicateurs sont transposés de l'analyse des réseaux sociaux présents dans les travaux en sociométrie.

La sociométrie se définit comme l'étude des groupes sociaux. L'analyse sociométrique s'intéresse aux interactions entre acteurs humains et les représente par des cartographies appelées réseaux et des indicateurs.

Plus d'un demi siècle après les premiers développements de la sociométrie, il serait très réducteur d'en rester au sociogramme de Moreno (1954). La sociométrie a connu des perfectionnements nombreux à travers les travaux dans le domaine de l'analyse des réseaux sociaux et a profité des perfectionnements dans le domaine de la visualisation de graphes, de la gestion de graphe de grande taille, autant de problèmes qui n'étaient même pas envisageables dans les perspectives technologiques de l'époque. Aussi ne souhaitons-nous pas nous réduire à la vision fondatrice de la sociométrie mais l'enrichir des travaux multidisciplinaires récents dans le domaine de l'analyse de réseau et de l'informatique.

Webométrie et sociométrie ont un suffixe commun mais curieusement, on ne retrouve pas de référence à la sociométrie dans les travaux en web-métrie.

---

[88] Selon l'algorithme de PageRank de *Google*, la pertinence d'une page dépend de la pertinence des pages qui la citent. Du fait de l'interaction entre les pages du web et du caractère petit monde du web, la pertinence d'une page ne peut pas être définie isolément des autres pages du web. Ce calcul de la pertinence est effectué de façon régulière par *Google* lors de la *Google* dance. A l'issue de la *Google* dance, chaque page web contenue dans l'index de *Google* se voit attribuer une valeur relationnelle. La pertinence d'une page web se définit donc en dehors du contexte de la requête.

La question naturelle qui vient à l'esprit est celle du rapprochement de la sociométrie et de la web-métrie. L'approche développée par l'analyse des réseaux sociaux est-elle transposable au cas du traitement de corpus web ? Certes dans les deux cas, on peut disposer de données en interaction. Toutefois, cette transposition ne va pas de soi pour deux raisons principales :

- Les quantités de données à traiter ne sont pas les mêmes. Les techniques sociométriques privilégient des petits réseaux alors que les analyses web-métriques ont souvent à gérer des réseaux de grandes tailles. Il est aisé de représenter un petit réseau de manière lisible et intelligible pour le décideur mais qu'en est-il lorsque ce réseau comporte plusieurs milliers voire plusieurs centaines de milliers de nœuds et arêtes ?
- Les techniques sociométriques travaillent sur des corpus où le lien entre deux sommets mesure un niveau d'interaction entre acteurs. Dans le cas de la web-métrie, l'interaction signifie l'existence d'un lien hypertexte entre deux pages web.

Passer de cartographies représentant des relations d'interaction entre des hommes à des cartographies mesurant des interactions hypertextuelles entre des pages web suppose un questionnement sur le sens du lien hypertexte. Le lien hypertexte est une Janus à double face.

- Dans sa logique de construction, le lien hypertexte obéit à différentes motivations qui peuvent être intrinsèques ou extrinsèques au web. Les motivations sont intrinsèques lorsque le lien hypertexte a pour objet la recherche d'un meilleur positionnement sur un moteur de recherche par exemple. Ce type de justification tend à se développer avec le développement du web marchand. Dans le cas où les motivations sont extrinsèques au web, le lien hypertexte est le témoin sur le plan virtuel d'une relation (hiérarchique, institutionnelle, informelle, de reconnaissance) existant ou ayant existé sur le plan réel. Nous considérerons comme Gérad Dubey (2001) que : « Parler du virtuel dans l'absolu revient […] à admettre la possibilité d'une réalité sans référence à ce qui existe ou a existé. Or une telle réalité *ex nihilo* est naturellement introuvable. Les relations qui se tissent sur le Web ont toutes pour modèle ou antécédent les relations sociales réelles ». Stuart et al. (2007) ont établi que l'existence de liens hypertextes au départ de

sites web universitaires reflétait une activité collaborative entre l'université et la cible du lien. Thelwall (2003) distingue 4 motivations au lien hypertexte : lien de navigation (la page est le point de départ d'une navigation qui permettra de trouver de l'information), liens de propriété (une page créée dans le cadre d'un consortium pointe vers les sites du consortium), liens sociaux (lien d'une page personnelle vers l'institution de rattachement), lien gratuit (cas de pages personnelles qui pointent vers une page où la personne a travaillé ou obtenu une qualification). Par ce lien hypertexte, le concepteur manifeste sa volonté implicite de reconstituer au niveau virtuel les relations qu'il a au niveau réel. La cartographie web-métrique relationnelle présente une carte difficile à interpréter puisque la signification et les logiques de liens hypertextuels sous-jacents sont multiples. De ce point de vue, l'analyse web-métrique manque de finesse et il faudrait pouvoir séparer automatiquement ces significations hypertextuelles. Mais la catégorisation automatique n'est pas facile.

- Dans la logique d'usage, toutes ces significations particulières du lien hypertexte disparaissent. Le lien hypertexte devient un tout, un objet à part entière qui va permettre, par son existence, une navigation dans un espace web. La logique web s'autonomise alors ; la représentation relationnelle de la sphère virtuelle n'est donc plus là pour éclairer la sphère réelle ; la présence dans la sphère virtuelle tend à s'imposer à elle seule. Le web va en effet être parcouru par les internautes qui vont suivre les chemins et les liens prévus à cet effet. Ce processus de navigation va permettre de passer d'un site à l'autre. Ainsi le réseau des interactions relationnelles est une construction macroscopique qui permet de mieux comprendre la façon dont l'internaute va déambuler. Les logiques à l'œuvre pour la construction de liens hypertextes nous intéressent moins car ce qui devient important c'est la façon dont ces chemins vont être empruntés.

Les analyses en termes de réseaux sociaux ont mis en évidence cinq formes de centralité principales : centralité de degré, de proximité, d'intermédiarité, de

flot et centralité au sens de l'information. Elles sont présentées par Degenne et Forsé (1994).

Nous en avons choisi deux : la centralité de degré et la centralité d'intermédiarité. Ce choix se justifie par le fait que ces deux indicateurs ont une interprétation assez intuitive, même si la centralité d'intermédiaire est plus complexe à calculer.

La centralité de degré considère qu'un sommet i est plus central qu'un sommet j si le sommet i a plus de sommets qui lui sont adjacents que le sommet j. Le calcul de la centralité de degré est aisé. Cet indicateur peut être décomposé en deux indicateurs appelés *Indegree* et *Outdegree*. *Indegree* correspond au nombre de liens entrants sur un sommet et outdegree au nombre de liens sortants d'un sommet. Cet indicateur donne une vision très locale de la centralité d'un sommet puisqu'il ne prend en compte que les sommets adjacents. Toutefois, dans la pratique, on observe que cet indicateur conduit à un classement très proche de celui d'autres indicateurs beaucoup plus sophistiqués.

Un sommet i est d'autant plus central au sens de l'intermédiarité qu'il appartient à un grand nombre de chemins géodésiques. Le chemin géodésique entre deux sommets j et k est le plus court chemin entre ces deux sommets. Lorsqu'un sommet se situe sur plusieurs géodésiques, cela signifie qu'il est le point de passage obligé entre de nombreux sommets du réseau. Cette situation peut s'apprécier également en termes de contrôle. Si i appartient au chemin géodésique entre j et k, cela signifie que i contrôle l'interaction entre j et k.

Pour calculer cet indicateur de centralité, il faut dénombrer l'ensemble des géodésiques correspondant aux paires de sommets (i, j) du réseau passant par chaque noeud. Ce calcul est gourmand. Nous ne l'avons pas implémenté.

Vers une facette centralité :

Pour obtenir ces deux indicateurs, il faut mettre en œuvre le processus suivant représenté dans la ***Figure 32***.

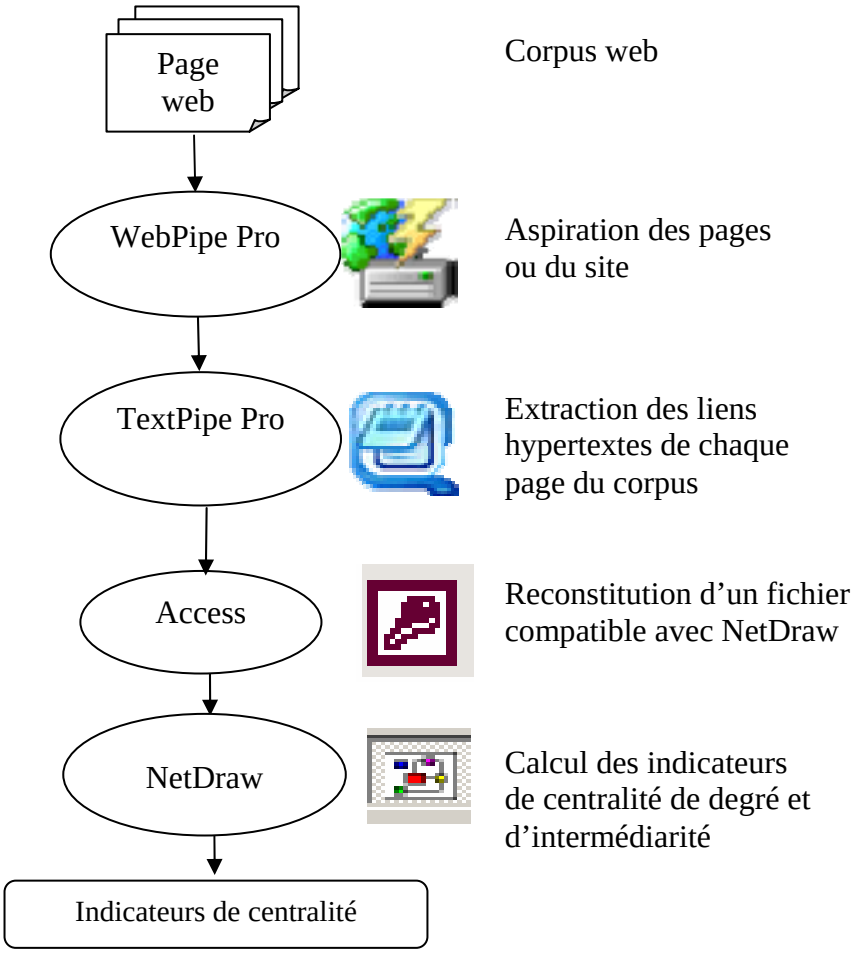


Figure 32 : chaîne de traitement de l'information pour déterminer les indicateurs de centralité

La chaîne de traitement semi automatique qui a été mise en œuvre mobilise les logiciels bureautiques traditionnels et des logiciels spécifiques (WebPipe Pro, TextPipe Pro, NetDraw).

Il faut au départ partir d'un corpus de pages web correspondant au sujet de la requête. Ces pages sont aspirées par le logiciel WebPipe Pro. Un retraitement sous le logiciel TextPipe pro permet d'extraire des pages web téléchargées les liens hypertextes qu'elles ont vers les autres pages du corpus[89]. Une fois que l'on dispose des pages web et des liens entre elles, il est possible de construire un fichier lisible sous NetDraw contenant la centralité de chaque sommet du réseau sous forme d'une valeur numérique. Il est alors aisé de classer les pages par valeur croissante et d'en déduire une facette dont les modalités peuvent être obtenues en découpant par exemple le corpus de pages en déciles.

[89] L'analyse peut être conduite au niveau des pages du corpus ou au niveau des sites web du corpus. Dans ce cas, le processus de téléchargement consiste à aspirer les pages web d'un domaine traitant du sujet de la requête

Nous avons implémenté cet indicateur. Aucune validation ne peut être faite car l'usager n'a pas les moyens d'apprécier le niveau de centralité d'une page web dans son contexte.
D'autres logiques d'indicateurs peuvent être envisagées pour mesurer la position relationnelle d'une page dans son contexte. Kleinberg (1999) définit à cet effet la notion de Hub et Authority. Les pages de type hub pointent sur des pages qui font autorité sur un sujet et des pages autorités sont citées par de bons hubs. Nous n'avons pas choisi d'implémenter ce type d'indicateur.

G. Classification par genre.

Ce type de facette est défini à partir de critères d'analyse stylométrique et de surface.

Dans une thèse récente sur le sujet, Elisabeth Sugar Baese (Sugar Baese - 2005) définit le genre d'un texte comme une taxonomie intégrant le style, la forme et le contenu d'un document. Cette taxonomie est orthogonale au sujet de la page web. Cela signifie qu'un même sujet peut être présenté sous forme de pages web de genres différents et qu'un genre peut permettre de présenter divers sujets. Quelques exemples de genres de pages web peuvent être donnés : pages de chat, page de magasine en ligne, article scientifique. Plusieurs auteurs (Santini - 2006, Rosso - 2005, Branca - 1999) se sont intéressés à une typologie des genres de pages web. Rousseau (2001) propose une typologie autour de 116 genres. Meyer zu Eissen et al. (2004) distinguent 8 genres (pages d'aide, article, page de forum, pages commerciales, portail privé, portail non privé, hub, page de téléchargement). Werlich (1976) distingue les pages de narration, de description, d'exposition, d'argumentation et d'instruction. Biber (1988) distingue les narrations personnelles engagées, les pages persuasives argumentatives, les pages conseils et les pages de type présentation technique. On s'aperçoit qu'en filigrane dans plusieurs de ces typologies, la nature subjective d'un texte est une dimension fondamentale autour de laquelle certains genres peuvent être décrits.

Dans ce travail, nous n'avons pas souhaité introduire de facette « genre ». En effet, par définition, chaque facette désigne une dimension orthogonale aux autres. Or, la facette genre peut être établie sur la base d'une analyse stylométrique, relationnelle, visuelle de la page. Certaines de ces dimensions sont déjà considérées comme des facettes. Si on veut respecter le principe d'orthogonalité des facettes, il n'est pas possible de considérer une facette genre qui serait une forme de métafacette construite à partir de facettes que nous avons considérées.

H. Analyse du trafic sur internet

On pourrait aussi distinguer les pages web selon le nombre de visiteurs qui ont visité cette page sur une période donnée. S'intéresser à cet indicateur d'usage, c'est reconnaître que la demande de l'usager contribue à valider la pertinence d'une information. Cette hypothèse est discutable dans la mesure où le trafic obtenu par une page web provient de plus en plus des efforts réalisés par les concepteurs de cette page pour obtenir un bon référencement dans les moteurs de recherche. En s'intéressant au nombre de visiteurs d'une page, on privilégie le professionnalisme du référenceur de la page plus que la qualité intrinsèque de la page. Les internautes privilégiant l'information contenue dans les premières réponses du moteur, il est à craindre qu'un tel indicateur d'usage valorise finalement une information bien établie et peu innovante et qu'il soit assez redondant avec les algorithmes actuels des moteurs de recherche que notre approche cherche précisément à contourner et à enrichir.

L'information sur les usages est disponible pour chaque site web et accessible pour le webmestre d'un site donné (fichier .Log). Il est par contre plus difficile d'obtenir ce type de données pour un site web dont on n'est pas propriétaire. Le moteur Alexa[90] fournit une estimation du trafic d'un site web. Alexa a créé une barre d'outils que plusieurs millions d'internautes ont installée sur leur navigateur. Cette barre d'outils permet à Alexa de récupérer, de façon transparente pour l'internaute, les sites qu'il a visités. Cette information permet de produire des statistiques d'usage qui sont vendues mais qui sont aussi disponibles gratuitement sous certaines conditions. En explorant le moteur de

[90] http://www.alexa.com

recherche d'Alexa, il est donc possible de connaître l'estimation du nombre de visiteurs d'un site donné, l'estimation du nombre moyen de pages d'un site visité. Un indicateur de synthèse appelé classement de trafic hiérarchise les sites web en prenant en compte la dimension du nombre de visites et du nombre de pages visitées. L'utilisation de cet outil pose un certain nombre de problèmes :

- Limite due au manque de granularité. Alexa fournit l'information relative au trafic du site web et non au trafic d'une page web.
- Représentativité des résultats. Les sites dont le rang de classement de trafic est supérieur à 100 000 ne reçoivent pas un nombre de visiteurs suffisant pour autoriser une interprétation statistique des résultats[91].

Pour ces raisons, nous avons choisi de ne pas implémenter cette facette.

[91] dixit Alexa http://www.alexa.com/site/help/traffic_learn_more#reach

CHAP 5 : IMPLEMENTATION DES FACETTES EN RI

Une fois les facettes identifiées, il reste à les intégrer dans l'outil de Recherche d'Information. Nous commencerons (A) par dresser une typologie des interfaces de navigation dans une classification par facettes. Ces applications s'inscrivent souvent dans des contextes métier. Nous définirons (B) les solutions techniques possibles et la solution retenue.

A. Typologie des interfaces de navigation

Les navigations par facettes ont surtout été implémentées dans des sites web marchands. La e-boutique à facettes présente un avantage par rapport à une boutique réelle. Dans une boutique réelle, on peut avoir un ou deux critères de classement dans un rayonnage, pas davantage : les vêtements sont par exemple classés par taille, couleur, marque ou prix. Dans une boutique virtuelle, on peut laisser au visiteur le choix des critères de recherche qu'il souhaite privilégier. La souplesse du web permet ainsi de fournir plusieurs portails d'accès aux contenus.
Nous allons donc, dans un premier temps, voir de quelles façons s'exprime la théorie des facettes sur le web marchand avant de voir les expériences d'interface à facettes qui ont été conduites dans des contextes de recherche d'information.

Une étude, conduite en 2003 par Adkisson (2005) sur 75 sites de e-commerce américains montre que 69 % des sites ont recours à une classification par facettes. Cette classification par facettes est implémentée soit dans la recherche avancée soit dans l'interface standard sollicitée par l'internaute lorsqu'il effectue une recherche d'information. Le *Tableau 47* présente la ventilation entre ces deux catégories.

Niveau d'apparition de la navigation par facette	%
Utilisation des facettes en recherche avancée uniquement	6%
Utilisation des facettes en recherche standard uniquement	77%
Utilisation des facettes en recherche standard et avancée	17%

Tableau 47 : façons dont est implantée la navigation par facettes

Un autre critère de distinction d'une navigation par facettes est le caractère plus ou moins dynamique des modalités des facettes lorsque l'internaute exprime son choix sur l'une d'entre elles. En 2006[92], l'interface du site www.epicurious.com fournissait un exemple de classification par facettes dans lequel il n'y avait pas de redéfinition contextuelle dynamique des effectifs par rubrique en fonction de la requête exprimée. La classification apparaîssait dans le menu avancé (*Figure 33*). Cette interface rappelle l'interface avancée des moteurs de recherche à la différence qu'elle est centrée sur le monde des recettes.

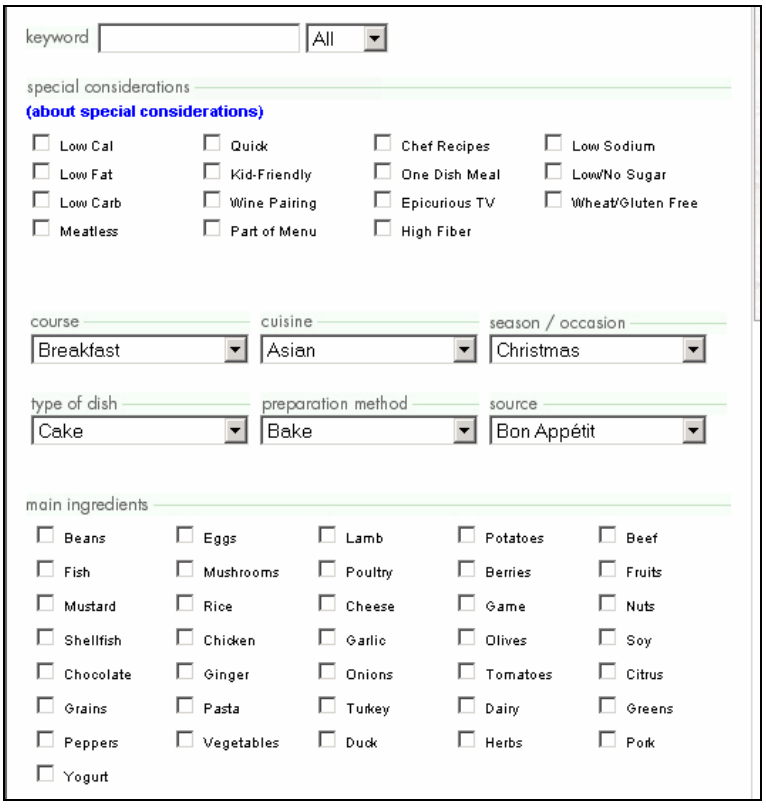


Figure 33 : interface par facette du site epicurious.fr

[92] l'interface du site www.epicurious.com a changé depuis 2006 et ne se présente plus sous forme de système à facette dans l'interface avancée.

Le système n'est donc pas très élaboré car le nombre de résultats ne se recalcule pas quand on coche une case. On est plus dans un système de recherche multicritères où on coche ce qui nous intéresse.

Le cas de www.Shopoon.fr présenté *Figure 34* est un exemple francophone de facettes définies de façon dynamique. Dans l'exemple choisi, on considère l'exploration d'un catalogue en ligne. Après avoir choisi un critère de sélection privilégié (vêtements pour enfants dans le cas de la *Figure 34*), on dispose de critères complémentaires permettant d'approfondir la recherche. Ces critères portent sur la marque, le prix, la couleur, le marchand. En affinant la recherche sur la base, par exemple, de la couleur (vêtement rose), les fréquences des modalités des autres facettes se recalculent en fonction du choix de l'internaute et restent affichées. Le résultat obtenu est présenté *Figure 35*. D'autres sites web marchands, tels amazon.fr pour sa partie vente de bijoux, utilisent ce type d'interface.
Notons que l'utilisation depuis peu de technologies web de type : AJAX permet la création de telles interfaces dynamiques.

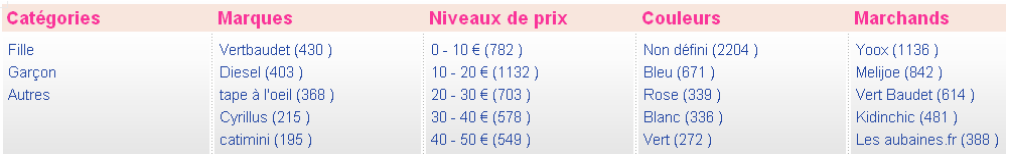


Figure 34 : interface par facettes – vêtements pour enfants

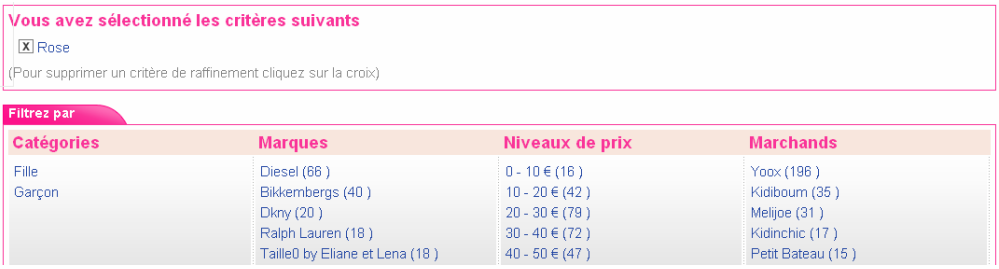


Figure 35 : Exemple d'interface à facettes avec redéfinition dynamique des contenus de chaque facette chez www.shopoon.fr

Brougton (2005) considère que si on reprend une définition rigoureuse des facettes, ce type de catégorisation ne correspond pas à quatre facettes

différentes mais à une seule facette entité ou personnalité qui se subdivise en quatre sous facettes.

Dans le domaine documentaire, l'Université d'Etat de Caroline du Nord innove en 2006 avec une interface de recherche d'information documentaire intégrant le principe de la théorie des facettes[93]. Lorsqu'il effectue une recherche d'information, l'internaute exprime sa requête. Une première liste de réponses lui est offerte mais il dispose également de moyens de navigation alternatifs où les documents correspondant à sa recherche sont classés par genre, auteur, langue, type de matériel, format, disponibilité. La *Figure 36* présente l'interface de cet outil d'exploration de la base documentaire : à droite figurent les résultats et à gauche les différentes facettes proposées.

Figure 36: interface à facettes de la bibliothèque de l'Université d'état de Caroline du Nord

[93] http://www.lib.ncsu.edu/searchcollection/

L'interface par facettes peut être implémentée d'une autre façon sous forme de curseurs.

Nous avons déjà donné plusieurs exemples de moteurs à curseurs à l'image du Mindset de Yahoo! qui permet de positionner les résultats d'une recherche sur un continuum commercial / recherche. Les moteurs de recherche commerciaux n'offrent généralement à la fois qu'un seul curseur mais rien n'empêche d'envisager et de concevoir un outil qui reprendrait plusieurs curseurs.

B. Implémentation : solution technique et solution retenue

Il existe différents outils logiciels permettant d'implémenter un moteur à facettes : Endeca, Flamenco projet de l'université de Berckley, FacetMap[94]. Nous allons décrire le fonctionnement de FacetMap. Cet outil permet de créer des fichiers textes qui sont ensuite chargés sur un site web et visibles sous une interface web. Le type de fichier est construit de façon assez intuitive. Considérons l'exemple fourni par FacetMap qui traite du domaine du vin. Chaque vin est défini autour de trois facettes : la variété du vin, son pays et son prix. Dans l'exemple ci-dessous, on observe que certaines facettes peuvent être organisées de façon hiérarchique : il y a plusieurs vins blancs, plusieurs vins français. La structure de ce fichier est fournie ci-dessous. Le fichier commence par décrire chacune des facettes :

[94] http://facetmap.com/demosetup/

```
%t
All varietals
0       red     Red Wines
# The above line has three elements, which must be separated by
# a 'Tab' character. First is the ID of the parent facet, which
# must have been defined in a previous line (but use "0" [zero]
# if this facet is a top-level facet in the taxonomy). Then,
# "red" is the ID given to this facet, and "Red Wines" is its
# full name. This goes on until a "%r" or another "%t" label is
# found.

Varietal
0       whi     White Wines
0       red     Red Wines
red     bar     Barbera
red     cab     Cabernet Sauvignon
red     cm      Cabernet/Merlot Blend
red     chi     Chianti
whi     rie     Riesling
whi     sau     Sauvignon/Fume Blanc
whi     sem     Semillon
whi     vio     Viognier
whi     wbo     White Bordeaux
%t
Region
0       aus     Australian
0       ita     Italian
0       fra     French
fra     als     Alsace
fra     bor     Bordeaux
ita     pie     piémont
%t
Price
0       low     Inexpensive (under $20)
0       mid     Mid-priced ($20-100)
0       high    Ultra-classy (over $100)
low     10      Under $10
low     20      $10-$20
mid     30      $20-$30
mid     40      $30-$40
%r
 Giacosa 1999 Barbera dAlba
bar pie 30

 Ca Rome 1999 La Gamberaja Barber
bar pie 40

 Ca Rome 1998 La Gamberaja Barbera d Alba
bar pie 30

 Ca' Bianca 1999 Barbera dAsti
bar pie 20

 Luigi Einaudi 1999 Barbera Piedmont
bar pie 30
...
```

Description de la facette variété du vin

Description de la facette pays

Description de la facette prix

Pour chaque vin, la première ligne décrit son nom et la seconde la modalité prise par chaque facette

Le site web de FacetMap propose une visionneuse intégrée qui fournit différents formats d'affichage ; Le format par défaut est présenté *Figure 37*.

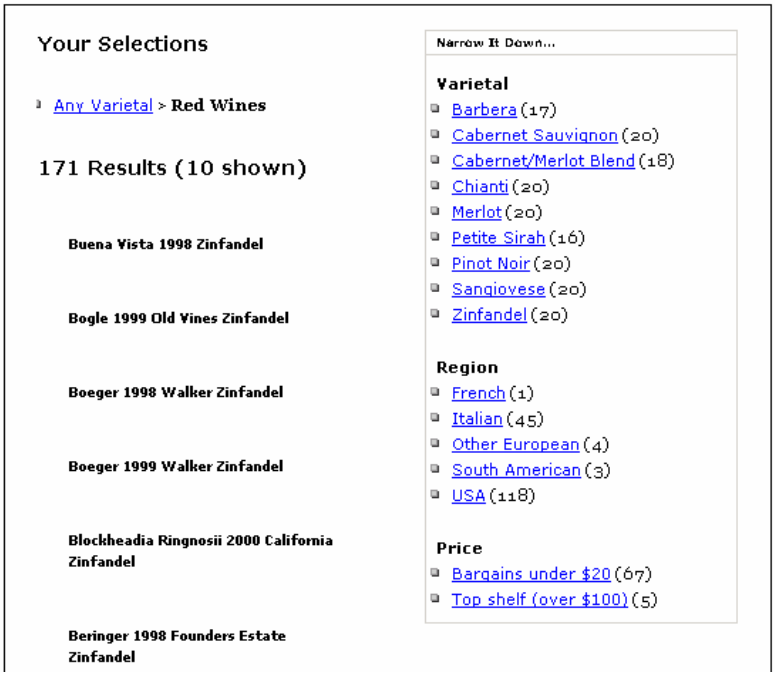


Figure 37 : visualisation de l'interface à facettes par défaut de FacetMap

La visionneuse peut également être téléchargée. Par contre dans ce cas, il faut être capable de générer des fichiers xml comportant les informations de la classification par facettes.
Dans le cadre de ce travail, une interface lisible sous FacetMap a été réalisée[95].

Une alternative consiste à développer un moteur de recherche dans lequel l'interface à facettes serait obtenue par l'intermédiaire de multiples curseurs. Cette alternative consisterait donc en une généralisation des initiatives prises par les outils de recherche.
L'interface se présenterait alors sous la forme d'une zone de saisie de la question de recherche et d'une liste de curseurs (deux curseurs sont représentés dans la *Figure 38*). L'internaute positionnerait les curseurs là où il le souhaite et pourrait ainsi choisir de donner un poids particulier à chaque facette. Dans l'exemple ci-dessous, on est à la recherche de pages négatives et subjectives et on accorde 3 fois plus de poids au caractère subjectif qu'au caractère négatif.

[95] http://facetmap.com/browse/websearch?v=rightnav.xsl&s=000000&n=0

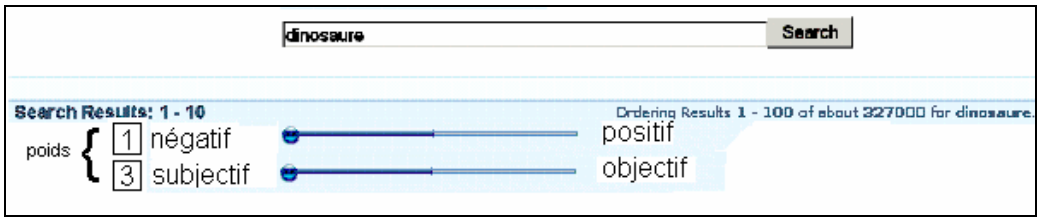


Figure 38 : exemple d'interface par facettes généralisant la notion de curseur

Lorsque l'internaute cliquerait sur le bouton search, la machine classerait les pages pour chaque critère et leur donnerait un rang. La machine calculerait ensuite, par combinaison linéaire, le poids de chaque document en prenant en compte le poids de chaque critère. Elle opérerait, après combinaison linéaire, un classement et restituerait les données dans le bon ordre.

Nous avons donc à disposition deux interfaces : une interface de menu contextuel dynamique permettant de présenter les modalités de chaque facette et une interface se présentant sous forme de curseurs. Nous avons choisi de représenter nos données grâce au menu contextuel dynamique et cela pour trois raisons.

- Le système par curseur s'avère particulièrement adapté à des variables où les objets étudiés peuvent être positionnés sur une échelle où figurent deux opposés. Certaines variables sont toutes indiquées pour faire l'objet de ce type de curseur (variable prix par exemple). Pour d'autres variables qualitatives, le curseur est moins adapté.
- Le second problème associé au recours aux curseurs est qu'ils ne se redéfinissent pas de façon dynamique en fonction du choix de tel ou tel élément. De ce point de vue, cette interface semble donc beaucoup moins puissante qu'une interface qui permet de redéfinir, en dynamique, le nombre de réponse de chaque modalité.
- Il n'existe pas de solution en libre accès permettant de visualiser une interface à curseurs alors qu'il en va différemment des interfaces dynamiques.

La phrase la plus stimulante à entendre en science, celle qui annonce de grandes découvertes, n'est pas "eurêka!" mais "tiens, c'est drôle..."
Isaac Asimov

CONCLUSION

Apports

Selon Ellis (1999), le génie de Ranganathan tient à la portabilité de ses idées dans d'autres cultures, d'autres univers technologiques, d'autres temps. L'analyse par facettes qui est conduite dans ce travail est très éloignée de celle du fondateur. Faut-il d'ailleurs encore parler de facettes ou le terme de multidimensionnalité (Pollity Y.- 2005) est-il plus approprié ? Nous avons conservé le terme de facettes pour deux raisons principales. La première est qu'il reste persistant dans la littérature anglosaxonne sur le sujet dont nous nous sommes nourri. La seconde raison est que le terme de facette, fournit une image éclairante : chaque facette constitue un « point de vue », un regard porté sur le monde.

L'originalité de ce travail par rapport aux perfectionnements récents apparus dans l'application au web de la théorie des facettes tient à la nature des facettes choisies. Il ne s'agit plus de décrire son besoin informationnel uniquement par une requête traduisant le contenu thématique du document recherché. Il s'agit aussi de le caractériser par des dimensions complémentaires (appelées facettes) qui ne renvoient pas seulement au contenu thématique des documents. Nous en avons identifié plusieurs et retenu cinq dans l'implémentation que nous avons proposée : le niveau de polarité d'une page web, le niveau de subjectivité d'une page web, le niveau d'accessibilité d'une page web, le niveau de lisibilité d'une page web et la centralité d'une page web dans son contexte. Chacune de ces dimensions a fait l'objet de développements théoriques et pratiques dans des domaines scientifiques d'appartenance, par exemple la linguistique computationnelle ou la psychologie cognitive. Notre objectif a consisté à aller chercher ces concepts et à étudier dans quelle mesure il étaient transposables à l'analyse de corpus web.

Ces nouvelles dimensions nous semblent en effet à même de résoudre un problème grandissant sur le marché de l'information numérique : une offre

d'information quantitativement massive et qualitativement très hétérogène est confrontée à une demande d'information émanant d'acteurs cherchant à satisfaire des besoins d'information très différents. Nous proposons alors une vision originale dans laquelle chaque facette va permettre à l'internaute d'affiner son besoin d'information et de le faire correspondre à la variété des réponses renvoyées par le moteur de recherche.

Nous proposons à l'internaute plusieurs regards sur une collection de documents. Il lui est offert le choix des regards ce qui lui permet de constituer une vue sur le monde correspondant à sa vision des choses. Cette approche permet donc une contextualisation fine en fonction du besoin de l'internaute.

Ce travail constitue une contribution au développement d'un moteur de recherche de type 2.0. Le web 2.0 sollicite l'internaute pour qu'il devienne acteur et non seulement utilisateur des ressources du web. Les logiques de web 2.0 appliquées à la recherche d'information existent aujourd'hui à travers les moteurs à Folksonomie. Il s'agit alors pour l'internaute acteur de qualifier des ressources et de rendre cette qualification visible à la communauté des autres internautes. Dans le travail que nous avons conduit, nous ne mobilisons pas l'utilisateur pour qu'il contribue à sa mesure et à son niveau à la construction de l'édifice d'un moteur de recherche à folksonomie. Nous souhaitons seulement lui « redonner la main » afin qu'il puisse exprimer ses préférences lors du processus de recherche d'information. Ces deux approches de type 2.0 ne sont d'ailleurs pas incompatibles. On pourrait demander aux internautes de qualifier les pages web qu'ils visiteraient selon des facettes de subjectivité, de lisibilité, de polarité, d'accessibilité. Cette qualification serait alors à la base de la détermination des facettes, l'internaute ayant encore le choix de combiner ces facettes dans son processus de recherche d'information. Cette évaluation humaine serait sans doute plus conforme à l'évaluation humaine qu'en font les internautes puisqu'elle émanerait d'eux. Toutefois, pour être significative, cette qualification doit être adoptée par un gros moteur à folksonomie.

Dans ce travail, nous considérons que la communication entre émetteur et récepteur ne se réduit pas au modèle de Shannon. Dumas (2005) prend la métaphore du modèle quantique de la lumière et distingue la dimension

granulatoire et ondulatoire de l'information. « La dimension granulatoire concerne tout ce qui est numérique, quantifiable, rationnel ou rationalisable ». La dimension ondulatoire s'intéresse « à ce qui est impalpable, de l'ordre de l'intuitif et des sentiments ». Nous considérons que cette frontière entre granule et onde n'est pas figée et qu'il est possible, de façon prudente, de formaliser certains éléments ondulatoires de manière granulatoire. Ainsi en va-t-il par exemple de la polarité d'un discours. Parler de polarité d'un discours, c'est s'attacher à la perception positive ou négative que va avoir l'usager d'un document. Cette perception repose sur le contenu mais va faire intervenir des éléments qualitatifs. La polarité d'un texte pour un internaute va dépendre de cette dimension ondulatoire au moins autant que de la dimension granulatoire. Si on ne prend en compte dans les indicateurs de pertinence des moteurs de recherche que de la dimension granulatoire (le contenu du document), on renonce à construire un indicateur de pertinence fidèle à la perception de l'usager. Si on veut, comme par mimétisme, construire un indicateur de pertinence de moteur de recherche fidèle à la perception de l'usager, alors il faut, avec tous les risques induits, s'intéresser à la dimension ondulatoire.

Analyse critique

Plusieurs critiques peuvent être adressées à ce travail.

- Le risque de « naturalisation ». Transformer des éléments aussi fins que le niveau de subjectivité, de lisibilité, de polarité d'un texte en indicateur automatique attribué par un ordinateur à l'issue de l'application d'un algorithme peut choquer le spécialiste de ces questions. Pour cette raison, nous avons souhaité, pour chaque facette, confronter les indicateurs automatiques aux jugements d'experts humains. Dans les différentes facettes sur lesquelles nous avons travaillé, nous avons observé, lors des expérimentations, que les juges humains sont d'accord entre eux sur le statut d'une page web dans 60% à 80% des cas. Ce relatif consensus nous invite alors à chercher à caractériser ce dénominateur commun (observé pour chaque facette) par un indicateur automatique. Notre approche ne consiste pas à réduire la complexité du réel à un indicateur automatique en prétendant rendre compte de tout. L'objectif est de réduire la complexité du monde à plusieurs indicateurs automatiques permettant de qualifier certaines dimensions d'un corpus documentaire et de

laisser ensuite l'internaute exprimer ses préférences vers telle ou telle facette ou combinaison de facettes. On voit donc que ce risque de naturalisation s'il peut être perçu au niveau de chaque facette, s'efface au niveau de l'interface globale. En effet, l'objectif final n'est pas de « mettre en boite » le réel mais de proposer une « boite à outils » qui permet à l'internaute de combiner entre eux plusieurs outils pour se créer sa vue sur le monde. Il y a donc un « lâcher prise » de la technique vers l'humain. L'opposition entre les sciences de la nature (sciences du généralisable) et sciences de la culture (sciences du particulier) n'a pas de prise sur ce travail qui dans un premier temps s'intéresse aux invariants pour mieux ouvrir la porte, dans un second temps, à l'expression du particulier.

- Le manque d'un travail sur les usages. Le travail que nous avons conduit débouche sur la réalisation d'un démonstrateur. Ce démonstrateur permet de visualiser, pour un corpus documentaire donné, (dans le domaine de l'énergie nucléaire), une interface qui met en jeu les différentes facettes que nous avons identifiées. Ce démonstrateur ne peut pas être utilisé pour faire une recherche d'information dans un contexte réel. Il a plus pour objectif de montrer ce que pourrait donner la technologie une fois implémentée. L'efficacité de l'interface de recherche par facettes n'est donc pas démontrée dans ce travail. La justification des facettes n'est pas davantage apportée par l'usage qu'auraient pu en faire les internautes. Les facettes ont été choisies par l'expertise et l'expérience. Ce démonstrateur est une étape nécessaire pour aller plus loin. Aller plus loin nécessite des moyens pour déployer les algorithmes[96] et gérer des bases de données géantes. Ce n'est que lorsque l'analyse par facettes sera déployée sur un moteur qu'elle pourra être testée par les internautes et donner ce retour d'usage. La conduite d'un travail sur les usages d'un tel dispositif est donc largement conditionnée par la recherche de moyens humains et financiers complémentaires. Par contre, rien n'oblige à implémenter toutes les facettes en même temps. La facette polarité d'une page web fait par exemple l'objet d'une transposition au Chinois Mandarin par un

[96] Le calcul des facettes pour chaque dimension ne s'effectuerait sans doute pas en temps réel mais lors de la constitution de la base de données.

laboratoire de l'école normale supérieure de Shanghai avec lequel nous sommes en partenariat. L'objectif est de construire un indicateur de polarité pour le chinois Mandarin qui sera utilisé dans des contextes d'intelligence économique et de veille d'image. La transposition au chinois ne représente pas seulement un changement de langue. La structure du Chinois fait que les mots ne sont pas séparés par des espaces comme en français. Avant de pouvoir calculer l'orientation positive ou négative d'un terme, il faut pouvoir séparer automatiquement chaque mot de la page. D'autre part, les outils logiciels sur lesquels nous avons travaillé pour récupérer les pages web et en extraire le texte raisonnent sur des pages en fichier texte. Le Chinois n'est pas géré par ces outils.

Perspectives et prolongements

Ce travail de recherche peut déboucher sur des perspectives intéressantes pour autant qu'on puisse passer d'une dimension d'un terrain de recherche individuel à un chantier de recherche collectif. Ce travail de recherche a permis de proposer une application originale et en même temps moderne de la théorie des facettes. Comme nous l'avons précisé plus haut, le passage du démonstrateur au prototype demande des moyens et oblige à s'inscrire dans une logique de réseau pour plusieurs raisons :

- Des raisons scientifiques : ce travail est un travail charnière entre l'informatique, les sciences de l'information et de la communication et d'autres disciplines. Aller pus loin dans ce travail suppose par exemple de s'associer avec des spécialistes en informatique.
- Les moyens financiers doivent être trouvés auprès de partenaires publics ou privés. Il s'agit alors d'identifier et de s'insérer dans des logiques de réseaux de financements de la recherche. De plus en plus, les moyens sont attribués à des entités travaillant en équipe.

Telle est sans doute une des responsabilités d'un Directeur de recherche. Ce travail suppose des qualités bien différentes de celles mobilisées par le chercheur qui fait sa recherche seul. Il recquiert une capacité à faire travailler en réseau des gens qui appartiennent à des univers différents, une aptitude à

monter des dossiers scientifiques, à convaincre pour mobiliser des moyens financiers, une connaissance fine des sources de financement à mobiliser et une stratégie d'influence et de lobbying pour accroître ses chances de succès. J'ai déjà initié de tels projets mobilisant des chercheurs provenant de disciplines diverses et s'associant dans le cadre d'un projet ayant eu un financement. Dernièrement, j'ai participé au montage d'un projet avec l'école normale supérieure de Shanghai. Nos partenaires chinois ont une réactivité extrême et sont capables d'apporter un premier financement très rapidement avant d'abonder si les résultats sont à la hauteur des espérances.

Dans un petit laboratoire de recherche, un chercheur doit être à la fois inventif pour trouver une idée originale, organisé pour transformer cette idée en recherche et connecté à des réseaux humains et financiers pour valoriser au mieux son projet. Cela demande des qualités nombreuses qu'il est difficile de retrouver chez la même personne. Ces étapes n'obéissent pas à la même logique temporelle (Lacroux et Nourry). L'activité de réseautage est très consommatrice en temps et en énergie ; l'activité inventive suppose des contextes très particuliers qu'il est difficile de reproduire et de formaliser ; la production de la recherche suppose un travail continu : ces trois exigences sont difficilement conciliables.

Comme nous l'avons précisé plus haut, cette contribution à la construction d'un moteur de recherche de type 2.0 est ouverte en ce qu'elle autorise la création de nouvelles facettes. Pour les moteurs de recherche, nous envisageons de travailler au déploiement de logiques de type web 3.0 qui s'intéresse au contenu textuel mais aussi sémiotique d'une page web. Nous réfléchissons à de nouvelles facettes qui décriraient une page web par des caractéristiques non sémantiques. Ces caractéristiques nouvelles pourraient mobiliser par exemple les dimensions colorimétriques et d'harmonie de la composition de la page web.

De façon plus générale, au delà de ce travail d'Habilitation à Diriger des Recherches, mon objectif est de continuer à m'investir dans le domaine des sciences de l'information et de la communication que je considère non comme

un « territoire à défendre » mais comme une source d'inspiration pour « créer des passerelles et conquérir la connaissance »[97].

[97] Selon les termes de Jacques Perriault, 16ième congres des SIC, première table ronde - 2008

BIBLIOGRAPHIE ET PUBLICATIONS DE L'AUTEUR

A. Bibliographie générale

Adkisson, H. P. (2005), « Web design practices: use of faceted classification », disponible en ligne : www.webdesignpractices.com/navigation/facets.html, consulté le 05/07/2008

Asch, S.E. (1946), Forming impressions of personality, *Journal of Abnormal and Social Psychology*, Vol. 41, pp. 258-90

Baeza-Yates, R., Saint-Jean, F., Castillo, C. (2002), "Web Structure, Dynamics and Page Quality", Proceedings of the *9th International Symposium on String Processing and Information Retrieval*, p.117-130, September 11-13, disponible en ligne : http://www.dcc.uchile.cl/~ccastill/papers/baeza04_web_dynamics_structure_page_quality.pdf, consulté le 07/02/2008

Barabási A.-L., Albert, R. (1999), Emergence of scaling in random networks, *Science*, 286, pp. 509-512, disponible en ligne : http://www.nd.edu/~networks/Publication%20Categories/03%20Journal%20Articles/Physics/EmergenceRandom_Science%20286,%20509-512%20(1999).pdf, consulté le 05/07/2008

Bar-Ilan, J. (2002), How Much Information Search Engines Disclose on the Links to a Web Page? – A Longitudinal Case Study of the 'Cybermetrics' Home Page, *Journal of Information Science*, vol. 28, no 6, pp. 455-466

Baroni, M., Vegnaduzzo, S. (2004). "Identifying subjective adjectives through web-based mutual information", In Ernst Buchberger, editor, Proceedings of *KONVENS 2004*, pages 17–24

Bast, H., Weber, I. (2006) "When you're lost for words: Faceted search with auto autocompletion", In *SIGIR'06 Workshop on Faceted Search* , pages 31–35, 2006.

Bates, M. (2002), "After the Dot-Bomb: Getting Web Information Retrieval Right This Time." FirstMonday 7 Juillet 2002, disponible en ligne : http://firstmonday.dk/issues/issue7_7/bates/index.html, consulté le 05/07/2008

Beaudouin, V., Fleury, S., Pasquier, M., Habert, B. et Licoppe, C. (2003), Décrire la Toile pour mieux comprendre les parcours, *Réseaux*, 116, pp. 19-51.

Belkin, N.J., Oddy, R.N. & Brooks, H.M. (1982) ASK for information retrieval: Parts 1 & 2, *Journal of Documentation* 38, 2/3 61-71, 145-164.

Björneborn, L. et Ingwersen, P. (2004), Toward a basic framework for Webometrics, *Journal of the American Society for Information Science and Technology* 55, No 14, 2004.

Bernard, F. & Joule, R.-V. (2004), Lien, sens et action : vers une communication "engageante" , *Communication & Organisation*, GREC/O, Université Michel de Montaigne, Bordeaux 3, 24, 347-362.

Bernard, F. & Joule, R.-V. (2005), Le pluralisme méthodologique en SIC à l'épreuve de la "communication engageante", *Questions de communication*, Presses Universitaires de Nancy, 7, 187-205.

Bestgen, Y., Fairon, C., Kevers, L. (2004), « Un baromètre affectif effectif : corpus de référence et méthode pour déterminer la valence affective de phrases », *JADT* 2004 : 7es Journées internationales d'Analyse statistique des Données Textuelles, disponible en ligne : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_016.pdf, consulté le 05/07/2008

Biber, D. (1988), *Variation across speech and writing*, Cambridge: Cambridge University Press

Boese, E. S. (2005), "Stereotyping the web: Genre classification of web documents", Master's thesis, Colorado State University, Fort Collins, CO, disponible en ligne : http://www.cs.colostate.edu/~boese/Research/masters.pdf, consulté le 05/07/2008

Bourque, G. (1989), « Des mesures de lisibilité », Communication présentée au 57e *Congrès de l'ACFAS*. Montréal

Branca-Rosoff, S. (1999), Types, modes et genres : entre langue et discours, *Langage et Société*, 87, pp. 5-24, disponible en ligne : http://www.cavi.univ-paris3.fr/ilpga/ED/dr/drsb/sb-pdf/intro-Branca-LS87.pdf, consulté le 05/07/2008

Briand, S, Tricaud, S, Michel, P. (2002) « Les limites de l'utilisation du coefficient Kappa », In: Actes des XXXIVèmes *Journées de Statistiques*; 2002 13-17 mai 2002; Bruxelles et Louvain-la-Neuve, Belgique; 2002. p. p. 301-2, disponible en ligne : http://www.stat.ucl.ac.be/jsbl2002/Kappamichel.pdf, consulté le 05/07/2008

Brin, S., Page, L. (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer networks*, 30, pp. 107-117, disponible en ligne : http://infolab.stanford.edu/pub/papers/*Google*.pdf, consulté le 07/02/2008

Broder, A. (2002), "A taxonomy of web search", *ACM SIGIR Forum*, Vol.36, No.2, pp. 3-10, disponible en ligne : http://www.acm.org/sigs/sigir/forum/F2002/broder.pdf, consulté le 05/07/2008

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J. (2000), Graph Structure in the Web, *Journal of Computer Networks*, 33(1-6),309-320, disponible en ligne :. http://www9.org/w9cdrom/160/160.html, consulté le 05/07/2008

Brody, T., Stamerjohanns, H., Vallières, F., Harnad, S., Gingras, Y., Oppenheim, C. (2004), "The effect of open access on citation impact", disponible en ligne :, http://users.ecs.soton.ac.uk/harnad/Temp/OA-TAadvantage.pdf, consulté le 05/07/2008

Broughton, V. (2005), The need for a faceted classification as the basis of all methods of information. retrieval,. *Aslib Proc*, 58(2), pp. 49–72

Casey, W., Navendu, G. et Shlomo, A. (2005), « Using Appraisal Taxonomies for Sentiment Analysis », disponible en ligne : http://lingcog.iit.edu/doc/appraisal_sentiment.pdf, consulté le 05/07/2008

Chakrabarti, S., Dom, B. , Gibson, D., Kleinberg, J., Kumar, S.R, Raghavan, P., Rajagopalan, S., Tomkins, A. (1999), Mining the link structure of the World Wide Web, *IEEE Computer*, August 1999

Chen, C. (2006), CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, *Journal of the American Society for Information Science and Technology*, 57(3), 359-377.

Chirita, P.A., Nejdl, W., Paiu, R., Kohlschütter, C. (2005), "Using ODP metadata to personalize search". *SIGIR 2005*: 178-185, disponible en ligne : http://www.l3s.de/~chirita/publications/chirita05using.pdf, consulté le 05/07/2008

Cho, J., Roy, S. (2004), « Impact of Search Engines on Page Popularity », 13th *international conference on WWW*, page 20-29 ACM Press, disponible en ligne : http://oak.cs.ucla.edu/~cho/papers/cho-bias.pdf, consulté le 05/07/2008

Cicchetti, D.V., Fleiss, J.L (1977), Comparaison of the null distributions of weighted Kappa and the Cordinal statistic, *Appl. Psychol. Meas*., 1977, N°1, pp. 195-201, disponible en ligne : http://apm.sagepub.com/cgi/reprint/1/2/195.pdf , consulté le 05/07/2008

Claypool, M., Le, P., Wased, M., Brown, D. (2001), « Implicit interest indicators », Proceedings on the I*nternational Conference on Intelligent User Interfaces*, January 14 - 17, 2001, Santa Fe, NM USA, Pages 33-40, disponible en ligne : http://web.cs.wpi.edu/~claypool/papers/iii/iii.pdf, consulté le 05/07/2008

Cohen, J. (1960), A coefficient of agreement for nominal scales., *Educ. Psychol. Meas*., 1960, 20, 27-46

Cohen, J. (1986), A coefficient of agreement for nominal scales , This weeks citation classic, *Current contents,* numéro 3, 20 Janvier 1986

Couzinet, V. (2006), « Les connaissances au regard des sciences de l'information et de la communication : sens et sujets dans l'inter-discipline ». In actes du colloque international *Semaine de la connaissance*, M. Harzallah, J. Charlet, N. Aussenac-Gilles (ed.) Université de Nantes 26-30 juin 2006. Vol. 3, p. 1-6.

Cowie, R. (2003), Describing the Emotional States Expressed in Speech, *Speech Comm*., vol. 40, April 2003, pp. 5–32, disponible en ligne : http://www1.cs.columbia.edu/~julia/papers/cowie00.pdf, consulté le 05/07/2008

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M. (2000), "'FEELTRACE': An instrument for recording perceived emotion in real time". In: Proceedings of the *ISCA Workshop on Speech and Emotion*, Northern Ireland (2000) 19–24, disponible en ligne : http://www.dfki.de/~schroed/articles/cowieetal2000.pdf, consulté le 05/07/2008

David, A. (2005), L'Intelligence Économique et les Systèmes d'Informations : Problématiques et approches de solutions, disponible en ligne : http://hal.inria.fr/inria-00000255/en/, consulté le 05/07/2008

De Landsheere, G. (1973) "Le test de closure" Labor, Bruxelles

Degenne, A. Forsé, M., (1994), *Les réseaux sociaux*, Editions Armand Colin

Deleuze, G. (1986), *Cours à Vincennes. Le pli, disponible en ligne :* www.webdeleuze.com, consulté le 05/07/2008 .

Denton, W, (2003), "How to Make a Faceted Classification and Put It On the Web", disponible en ligne : http://www.miskatonic.org/library/facet-web-howto.html, consulté le 05/07/2008

Diaz, A. (2005), "Through the Google Goggles: Sociopolitical Bias in Search Engine Design". Thesis, Stanford University, disponible en ligne : http://epl.scu.edu/~stsvalues/readings/Diaz_thesis_final.pdf, consulté le 05/07/2008

Dubey, G. (2001), *Le lien social à l'ère du virtuel*, Paris, PUF.

Dumas, P. (2005), le distic et le rayonnement des cultures, Colloque *Culture des organisations et Distic*, Nice, les 8 et 9 Décembre 2005

Ekman, P. et Keltner, D. (1997), Universal Facial Expressions of Emotion: An Old Controversy and New Findings. In U. Segerstrole and P. Moln´ar, editors, Nonverbal Communication: Where Nature Meets Culture, pages 27–46. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, disponible en ligne : http://www2.dfki.de/~schroed/articles/cowieetal2000.pdf, consulté le 05/07/2008

Ellis, D, Vasconcelos, A. (1999), "Ranganathan and the Net: using facet analysis to search and organize the World Wide Web", *Aslib Proceedings*, 51(1) (Janvier 1999). P 3-10

Esuli A and Sebastiani F (2005), "Determining the semantic orientation of terms through gloss classification" In Proceedings of *CIKM-05*, 14th ACM International Conference on Information and Knowledge Management, Bremen, DE. Forthcoming, disponible en ligne : http://tcc.itc.it/projects/ontotext/Publications/CIKM05-short.pdf, consulté le 05/07/2008

Farradane, J. (1957) "Summury of the discussions", Proceedings of the *international conference on classification for information retrieval*, Dorking England, Mai 1957, P.91-108

Faucompré, P. (1997), « la mise en correspondance automatique de banques de donnees bibliographiques scientifiques et techniques à l'aide de la classification internationale des brevets contribution au rapprochement de la science et de la technologie », thèse soutenue le 13/10/1997, disponible en ligne : http://quoniam.univ-tln.fr/theses/faucompre/faucompre.pdf, consulté le 05/07/2008

Feinberg, M. (2007), "Beyond retrieval: A proposal to expand the design space of classification". Proceedings of the North American Symposium on Knowledge Organization. Vol. 1, disponible en ligne : http://dlist.sir.arizona.edu/1892, consulté le 05/07/2008

Fenelon, J.P., (1981), Q*u'est-ce que l'Analyse des données ?*, Paris, Lefonen, 1981. - 311 p

Fermanian, J. (1984), Mesure de l'accord entre deux juges. Cas qualitatif, *Rev. Epidém. et Santé Publ*., 1984, 32, 140-147.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press

Fetterly D. Manasse M., Najork M., Wiener JL. (2003), "A large-scale study of the evolution of web pages". In Proceedings of the *Twelfth WWW Conference*, Budapest, Hungary, disponible en ligne : http://www2003.org/cdrom/papers/refereed/p097/P97%20sources/p97-fetterly.html, consulté le 05/07/2008

Flesh, R. (1948), A new readibility yardstick, *Journal of Applied Psychology*, 32, 221233.

Fogg, B. (2003). Motivating, Influencing, and Persuading Users. In J. Jacko & A. Sears (Eds.), *The Handbook of Human-Computer Interaction* (pp. 359-370). Mahwah, NJ: Lawrence Erlbaum Associates.

Foskett, D.J. (1964), The Construction of a Faceted Classification for a Special Subject, *Science Humanism and libraries*, Crosby Lockwood, Londres, pp. 867-888

Furnas, G. W. (1986). *Generalized Fisheye Views*. CHI.

Garfield, E. (1984), A Tribute to S. R. Ranganathan, the father of Indian Library Science. Life and Works, *Current Contents*, #6, p. 5-12, February 6, 1984

Garfield, E. (1957) Summury of the discussions, Proceedings of the *international conference on classification for information retrieval*, Dorking England, Mai 1957, P.91-108

Gélinas-Chebat, C., Préfontaine, C., Lecavallier, J., Chebat, J.-C. (1993). " Lisibilité - Intelligibilité de documents d'information ". In Le projet SATO-CALIBRAGE, disponible en ligne :http://corpus.ato.uqam.ca/sato/publications/bibliographie/C3lisib.htm, consulté le 07/02/2008

GélinasChebat, C., Macot, M., Préfontaine, C., et Daoust, F. (1991), *La lisibilité de documents d'information du ministère de la Main d'oeuvre, de la Sécurité du revenu et de la Formation professionnelle*, Avis professionnel présenté au ministère de la Main d'oeuvre, de la Sécurité du revenu et de la Formation professionnelle, Gouvernement du Québec, 50 p

General Inquier, (2000), disponible en ligne : http://www.wjh.harvard.edu/~inquirer/3JMoreInfo.html, consulté le 05/07/2008

Glassel, A. (1998). Was Ranganathan a Yahoo !?, University of Wisconsin, Internet Scout Project, End User's Corner, disponible en ligne : http://scout.cs.wisc.edu/toolkit/enduser/archive/1998/euc-9803, consulté le 07/02/2008

Godby, J., Stuler, J. (2001), "The Library of Congress Classification as a knowledge base for automatic subject categorization." Presented at the *IFLA* Preconference, "Subject Retrieval in a Networked Environment," Dublin, Ohio, August 2001

Goffman, E. (1981), Response cries. In Goffman, E. Forms of Talk. 78-122. Oxford: Blackwell.

Goldstein, J., Carbonell, J. (1998), "Summarization: (1) using MMR for diversity - based reranking and (2) evaluating summaries, Annual Meeting of the ACL", Proceedings of a workshop on held at Baltimore, Maryland: October 13-15

Goria, S. (2006), « l'expression du problème dans la recherche d'informations : application à un contexte d'intermédiation territoriale », doctorat, sous la Direction d'Amos David, Université de Nancy 2.

Gunning, R. (1952), *The technique of clear writing*, New York: McGrawHill.

Harter, S. (1992), Psychological relevance and information science, *Journal of the American Society for Information Science*, 43(9), pp. 602-615.

Hatzivassiloglou, V., McKeown, K.R. (1997), "Predicting the semantic orientation of adjectives". In Proceedings of *ACL-97*, 35th Annual Meeting of the Association for Computational Linguistics, pages 174–181, Madrid, disponible en ligne : http://acl.ldc.upenn.edu/P/P97/P97-1023.pdf, consulté le 05/07/2008

Heine, M.H. (2000b). "Reassessing and Extending the Precision and Recall Concepts," In www.ewic.org.uk/ewic. Revised version of "Time to dump 'P and R'?" Proceedings of the *MIRA '99*: Final MIRA Conference on Information Retrieval Evaluation, Glasgow, 14-16 April 1999: 61-74, disponible en ligne : http://www.bcs.org/upload/pdf/ewic_mi99_paper4.pdf, consulté le 05/07/2008

Henry, G. (1975), *Comment mesurer la lisibilité*, Paris, Fernand Nathan, Editions Labor, 176p.

Herlocker, J., Konstan, J., Terveen, L., Riedl, J. (2004), Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22, ACM Press, 5-53. disponible en ligne : http://ectrl.itc.it/home/laboratory/meeting/download/p5-l_herlocker.pdf, consulté le 05/07/2008

Hindman, M. Tsioutsiouliklis, K. et Johnson, J. A. (2003), "Googlearchy: How a few heavily-linked sites dominate politics online", http://www.johnkeane.net/pdf_docs/teaching_sources/*Google*/*Google*.pdf, consulté le 07/02/2008

Hjørland, B. (1998), The classification of psychology: a case study in the classification of a knowledge field, *Knowledge Organization*, 25 (4): 162–201.

Holstrom, J E. (1957) Summury of the discussions, Proceedings of the *international conference on classification for information retrieval*, Dorking England, Mai 1957, P.91-108

Ingwersen, P. (1992), *Information Retrieval Interaction*. Taylor Graham, London, 1992, disponible en ligne : http://www.db.dk/pi/iri/files/Ingwersen_IRI.pdf, consulté le 07/02/2008

Joule, R.-V. & Beauvois, J.-L. (1987), *Petit traité de manipulation à l'usage des honnêtes gens*, Presses Universitaires de Grenoble, Grenoble, 2002.

Joule, R.-V. & Beauvois, J.-L. (1998), *La soumission librement consentie*, Paris, Presses Universitaires de France.

Juillet, A. (2006) « Référentiel de formation en intelligence économique », Sectérariat Général de la Défense Nationale, disponible en ligne : http://www.ceram.edu/v3/R7_modules/R7-5_contenu/R7-5-1_type1/R7-5-1-0_media/R7-5-1-0-1_contenu_html/ACF141F.PDF, consulté le 05/07/2008

Kamps, J., Marx, M., Mokken, R. J., Rijke, M. D. (2004), "Using WordNet to measure semantic orientation of adjectives". In Proceedings of *LREC-04*, 4th *International Conference on Language Resources and Evaluation*, volume IV, pages 1115–1118, Lisbon, PT, disponible en ligne : http://staff.science.uva.nl/~mdr/Publications/Files/lrec2004_wordnet.pdf, consulté le 05/07/2008

Kandel, Moles, A. (1958), Application de l'indice de Flesch à la langue française, *Cahiers d'Etudes de Radio-Télévision*, n° 19, pp. 252-274, Paris

Kim, K., Sei-Ching, J.S., Soo-Jin, P., Xiaohua, Z., Jom, P. (2006), "Facet Analyses of Categories Used in Web Directories: A Comparative Study", *IFLA* 2006, Séoul

Kincaid, J. P.; Fishburne, R. P., Jr.; Rogers, R. L.; and Chissom, B. S. (1975), "Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel", Research Branch Report 8-75, Millington, TN: Naval Technical Training, U. S. Naval Air Station

Kleinberg, J.M. (1999), Authoritative sources in a hyperlinked environment, *Journal of the ACM*, vol. (46/5), pp.604-632, disponible en ligne : http://www.cs.cornell.edu/home/kleinber/auth.pdf, consulté le 05/07/2008

Koch, T., Day, M., Brümmer, A., Hiom, D., Peereboom, M., Poulter, A., Worsfold, E. (1997), Specification for resource description methods Part 3: The role of classification schemes in Internet resource description and discovery. Work Package 3 of Telematics for Research project DESIRE - Development of a European Service for Information on Research and Education, février 1997, disponible en ligne : http://www.ukoln.ac.uk/metadata/desire/classification/classification.pdf, consulté le 05/07/2008

Kuronen, T., Pekkarinen, P. (1999), Ranganathan revisited: a review article, *Journal of Librarianship and Information Science*, 31; pp.45-48,

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, SS, Martimbeau, N., Elwell, B. (2003), The NASA Astrophysics Data System: Sociology, Bibliometrics, and Impact, disponible en ligne : http://cfa-www.harvard.edu/~kurtz/jasist-submitted.pdf, consulté le 05/07/2008

Kurtz, M.J. (2004), Restrictive access policies cut readership of electronic research journals articles by a factor of two., disponible en ligne : http://opcit.eprints.org/feb19oa/kurtz.pdf, consulté le 05/07/2008

Kyle, B. (1957) Summury of the discussions, Proceedings of the *international conference on classification for information retrieval*, Dorking England, Mai 1957, P.91-108

La Barre, K. (2004). "Weaving webs of significance: The Classification Research Study Group in the United States and Canada". In B. Rayward & M. E. Bowden (Eds.), Proceedings of the Second *Annual Conference on the History and Heritage of Scientific and Technical Information* Systems, 15–17 November, 2002. Medford, NJ: Information Today

Labasse, B. (1999), La lisibilité rédactionnelle: fondements et perspectives, *Communication et Langages*

Lacey, A. (2005), "A Simple Probabilistic Approach to Ranking Documents by Sentiment", Proceedings of the Conference Class of 2005 *Senior Conference on Natural Language Processing*, Spring 2005 Swarthmore College Swarthmore, Pennsylvania, USA, disponible en ligne : http://www.cs.swarthmore.edu/~richardw/cs97/papers/01_Paper.pdf, consulté le 05/07/2008

Lacroux, F. et Nourry, L. (1997). « Temps et rythmes de la stratégie. » ; Actes de la Conférence de *l'Association Internationale de Management Stratégique*, Montréal

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Landis, J.R., Koch, G.G. (1977), The Measurement of Observer Agreement for Categorical Data, *Biometrics*, 33, pp. 159-174.

Lavrenko V. (2004),." A Generative Theory of Relevance". Thèse de doctorat, University of Massachusetts, Amherst, MA, disponible en ligne : http://ciir.cs.umass.edu/~lavrenko/thesis.pdf, consulté le 07/02/2008

Lawrence, S. (2001), Free online availability substantially increases a paper's impact. Nature (Web Debates), (version éditée dans *Nature* 411, 521, disponible en ligne : http://www.nature.com/nature/debates/e-access/Articles/lawrence.html, consulté le 05/07/2008

Kahneman, D., Slovic, P. et Tversky, A. (1982), *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge University Press.

Kessler, M., (1963), Bibliographic couplic beetween scientific papers, *American Documentation* 14, No 1.

Kislin, P. (2007), « modélisation du problème informationnel du veilleur dans la démarche d'intelligence économique », doctorat, sous la Direction d'Amos David, Université de Nancy 2.

Macedo-Rouet, M., Rouet, JF., Zampa, V., Bouin, E. (2008), « l'information Internet : le jugement de crédibilité des usagers » communication, *congrès SFSIC 2008*

Mai, J.E. (2004), classification of the web : challenges and inquiries, *Knowledge Organization*, 31, N° 2, , pp. 92-97

Mai, J.E. (2004b), classification in context : relativity, reality and representation, *Knowledge Organization*, 31, N° 1, pp. 39-48

Maniez, J. (1999), Des classifications aux thésaurus : du bon usage des facettes, *Documentaliste- Sciences de l'information*, vol. 36, n° 4-5, pp. 249-262.

Mesnager, J. « Pour une étude de la difficulté des textes : la lisibilité revisitée », choses lues, disponible en ligne : http://www.etab.ac-caen.fr/centre-ph-lucas/carep/fichier/mesnager1.pdf, consulté le 05/07/2008

Meyer zu Eissen, S. , Stein, B. (2004) "Genre classification of web pages". In the Proceedings of the 27th *German Conference on Artificial Intelligence* (KI-2004), Ulm, Germany, September 20-24 2004, disponible en ligne : http://www.uni-weimar.de/medien/webis/publications/downloads/stein_2004c.pdf, consulté le 07/02/2008

Mizzaro (1998), How many relevances in information retrieval? *Interacting With Computers*, 10(3), 1998, 305-322, disponible en ligne : http://www.dimi.uniud.it/mizzaro/research/papers/IwC.pdf, consulté le 05/07/2008

Monnoyer-Smith, L. (2008), Pour une épistémologie complexe des SIC, *16 ième congrès des SIC*, Compiègne, Juin 2008

Moreno, (1954) *Fondements de la sociométrie*, PUF Paris 1954

Morin, E. (1982), *Science avec conscience*, Fayard, 1982.

Morin, E. (1999), *l'intelligence de la complexité*, l'Harmattan, Paris, 1999

Morin, E., Le Moigne, J.-L. (2007), « Intelligence de la complexité, Epistémologie et Pragmatique », Actes du *colloque de Cerisy*, Paris, Éd. de l'Aube

Morris, S.A., Yen, G., Wu, Z., Asnake, B. (2003), Timeline visualization of research fronts, *Journal of the American Society for Information Science and Technology*, 55(5), pp.413–422.

Mukras, R. (2004), "A Comparison Of Machine Learning Techniques Applied To Sentiment Classification", master thesis, disponible en ligne : http://www.comp.rgu.ac.uk/staff/ram/publications/papers/mukras04comparison.pdf, consulté le 05/07/2008

Nichols, D. M. (1997), "Implicit Rating and Filtering", In Proceedings of the *5th DELOS Workshop on Filtering and Collaborative Filtering*, Budapest, Hungary, 10-12 November 1997, ERCIM, 31-36, disponible en ligne : http://www.comp.lancs.ac.uk/computing/research/cseg/projects/ariadne/docs/delos5.pdf, consulté le 05/07/2008

Ntouloas, A., Cho, J., Olston, B. (2004), "What's New on the Web? The Evolution of the Web from a Search Engine Perspective", *WWW2004*, May 17–22, 2004, New York, New York, USA, page 1

Odlyzko, A. (2002), The rapid evolution of scholarly communication. *Learned Publishing* 15(1), 7-19.

Othon, W., Pan, W., Song, S. (2004) "examination of bias in search engine results using webes", final report hiver, disponible en ligne : http://dcm.cl.uh.edu/capf4gp5/pdf/fall2004_gp5_final_report.doc, consulté le 05/07/2008

Otlet, Paul. (1934), *Traité de documentation : le livre sur le livre : Théorie et pratique*. Brussels, éditiones Mundaneum

Pang, B., Lee L., Vaithyanathan, S. (2002), "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proceedings of *EMNLP* 2002.

Pang, B., Lee, L. (2004), "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Proceedings of *ACL* 2004.

Pang, B., Lee, L. (2005), "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales", Proceedings of *ACL* 2005.

Panteado, R. (2006), « Création de systèmes d'intelligence dans une organisation de recherche et développement avec la scientométrie et la médiamétrie », doctorat en sciences de l'information et de la communication, université du Sud Toulon Var, dirigé par Luc Quoniam, Septembre 2006

Persson, O. (1994), "The intellectual base and research fronts of JASIS 1986–1990", *Journal of the American Society for Information Science*, 45(1), pp.31–38.

Pierret, J. D. (2006), « Méthodologie et structuration d'un outil de découverte de connaissances base sur la littérature biomédicale : une application basée sur l'exploitation du Mesh », doctorat en sciences de l'information et de la communication, université du Sud Toulon Var, dirigé par Luc Quoniam, Février 2006

Pinczon du Sel, P. (2006), « Etat des lieux des résultats d'une recherche d'information simultanée sur le moteur de recherche Google. », Actes du colloque *ACSI/CAIS*, Université York, Toronto, Canada, Juin 2006. En ligne : http://archivesic.ccsd.cnrs.fr/sic_00001755, consulté le 05/07/2008

Pollity, Y., Henneron, G., Palermiti, R. (2005), *l'organisation des connaissances approches conceptuelles*, l'Harmattan, 2005

Quoniam, L. (1996), « Les productions scientifiques en bibliométrie et dossiers de travaux », Mémoire d'habilitation en sciences de l'information et de la communication, Université d'Aix- Marseille III, 1996.

Redish, J. C., Seizer, J. (1985). The place of readability formulas in technical communication, *Technical Communication*, 32(4), pp. 46-52.

Reymond, D., (2007), « Dynamique informationnelle d'une ressource Web : apport sémantique de la taxinomie. Etude webométrique des sites des universités françaises « , doctorat en sciences de l'information et de la communication, Université de Bordeaux, dirigé par Roland Ducasse, Décembre 2007

Ranganathan, S.R. (1933), *Colon Classification.* Madras, India: Madras Library Association

Ranganathan, S. R. (1967) *Prolegomena to Library Classification.* Asia Publishing House (New York), disponible en ligne : http://dlist.sir.arizona.edu/1151/, consulté le 05/07/2008

Richaudeau, F. (1993), *Manuel de typographie et de mise en page*, Paris, Retz.

Rose, D.E., Levinson, D. (2004), "Understanding user goals in Web search". In Proceedings of the *Thirteenth Int'l. World Wide Web Conf.*, 2004, disponible en ligne : http://www2004.org/proceedings/docs/1p13.pdf, consulté le 07/02/2008

Rosso, M. (2005), "Using genre to improve web search", Doctor of Philosophy in the School of Information and Library Science, University of North Carolina at Chapel Hill, disponible en ligne : http://ils.unc.edu/~rossm/Rosso_dissertation.pdf, consulté le 05/07/2008

Rousset, C., Benslimane, D., Arara, A., Vangenot, C. (2005), Contribution à l'interopérabilité dans les ontologies : la multi représentation », in *l'organisation des connaissances : approches conceptuelles*, l'Harmattan

Santini, M. (2006), "Interpreting Genre Evolution on the Web", *EACL 2006* Workshop: NEW TEXT - Wikis and blogs and other dynamic text sources, disponible en ligne : http://www.sics.se/jussi/newtext/working_notes/06_santini.pdf, consulté le 05/07/2008

Sarkar, M., Brown M. H. (1992). *Graphical Fisheye Views of Graphs*. CHI.

Schroder, M. (2004). "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions". In Proc. *Workshop on Affective Dialogue Systems*, pages 209–220, Kloster Irsee, Germany, disponible en ligne : http://www2.dfki.de/~schroed/articles/schroeder2004.pdf, consulté le 05/07/2008

Semin, G. R., Fiedler, K. (1988), The cognitive functions of linguistic categories in describing persons: social cognition and language, *Journal of personality and social psychology*, vol. 54, no4, pp. 558-568

Shah, L., Kumar, S. (2006), "Uniform Form Divisions (Common Isolates) for Digital Environment, A Proposal", *world library and information congress*: 72nd IFLA general conference and council, 20-24 August 2006, Seoul, Korea, .

Shannon, C.E. (1948), A Mathematical Theory of Communication, *Bell System Technical Journal*, 27, pp. 379–423 & 623–656, July & October, disponible en ligne : http://en.wikipedia.org/wiki/Special:BookSources/0252725484, consulté le 07/02/2008

Simmonot, B. (2002), « De la pertinence à l'utilité en recherche d'information : le cas du Web », Recherches récentes en Sciences de l'information - convergences et dynamiques, *actes du colloque international MICS-LERASS*, 21-22 mars 2002, Toulouse ; ADBS Éditions, collection Sciences de l'information, série Recherches et Documents, Paris, 2002, pp. 393-410, disponible en ligne : http://archivesic.ccsd.cnrs.fr/docs/00/06/26/04/PDF/sic_00001410.pdf, consulté le 05/07/2008

Simon, H. (1982) *Models of bounded rationality: Behavioral economics and business organization (Vol. 1 et 2)*, The MIT Press.

Small, H. (1973), Co-citations in the scientific literature : a new measure of the relationship beetween two documents, *Journal of the American Society for Information Science*, No 4, disponible en ligne : http://www.garfield.library.upenn.edu/essays/v2p028y1974-76.pdf, consulté le 05/07/2008

Solla Price, D. J. (1965), Network of scientific papers. The patterns of bibliographic references indicates the nature of scientific research front; *Science*, 149, pp. 510-515 mis dans la biblio lu bofpricenetworks1965

Spiteri, L. (1998), A Simplified Model for Facet Analysis, *Canadian Journal of Information and Library Science*, v23, pp. 1-30,

Stamatatos, E., N. Fakotakis, Kokkinakis, G. (2000), "Text Genre Detection Using Common Word Frequencies". In the Proc. of the 18 th *International Confernece on Computational Linguistics* (COLING2000)

Stuart, D., Thewall, M., Harries, G. (2007), UK academic web links and collaboration – an exploratory study, *Journal of Information science*, 2007- 33, 231

Swanson, D.R. (1986), Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspectives in Biology and Medicine*. Vol. 30, n°1, p. 7-18.

Swanson, D.R. (1993), Intervening in the life cycles of scientific knowledge, *Library Trends*. Vol. 41, n°4, p. 606-631.

Taher, H., Haveliwala, (2003), "Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search". *IEEE Trans. Knowl. Data Eng.*, 15(4):784--796, 2003, disponible en ligne : http://www-cs-students.stanford.edu/~taherh/papers/topic-sensitive-PageRank.pdf, consulté le 07/02/2008

Tan, B., Shen, X., Zhai, C. (2006), "Mining Long-term Search History to Improve Search Accuracy", Proceedings of 2006 *ACM Conference on Knowledge Discovery and Data Mining* (SIGKDD'2006), pages 718-723, disponible en ligne : http://sifaka.cs.uiuc.edu/czhai/pub/kdd06-pers.pdf, consulté le 05/07/2008

Taylor, R.S. (1968),Question-Negotiation and Information Seeking in Libraries, *college and Research Libraries*, 29

Taylor, A.G. (1999) *The Organization of Information.* Englewood, Colorado: Libraries Unlimited, 1999.

Teevan, J., Dumais, S. T. and Horvitz, E. (2005). "Beyond the commons: Investigating the value of personalizing Web search". *In Proceedings of the Workshop on New Technologies for Personalized Information Access* (PIA), disponible en ligne : http://research.microsoft.com/~sdumais/PIA2005-final.pdf, consulté le 05/07/2008

Thelwall, M. (2003), What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation, *Information research*, 8(3), disponible en ligne : http://informationr.net/ir/8-3/paper151.html, consulté le 05/07/2008

Travis, W. (2006), The strict faceted classification model, disponible en ligne : http://facetmap.com/pub/strict_faceted_classification.pdf, consulté le 05/07/2008

Trubert-Ouvrard T. (2001), « Adjectif antéposé ou postposé au nom : argumenter et convaincre dans le discours électoral », Actes *du Colloque sur les discours des élections municipales françaises de mars 2001* en Sorbonne le 13 juin 2001, disponible en ligne : http://www.seinan-gu.ac.jp/~trubert/adjectif4.html, consulté le 05/07/2008

Turney, P.D., Littman, M L., (2002), Un-supervised learning of semantic orientation from a hundred-billion-word corpus, Technical Report EGB-1094, National Research Council Canada, disponible en ligne : http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-44929.pdf, consulté le 05/07/2008

Tversky, A., Kahneman D. (1974), « Judgment Under Uncertainty: Heuristics and Biases », *Science*, 185, 1124-1131

Uichin, L. Zhenyu, L., Junghoo, C. (2005), "Automatic identification of user goals in Web search". In Proceedings of the *14th International World Wide Web Conference* (WWW) '05, Chiba, Japan, disponible en ligne : http://www2005.org/cdrom/docs/p391.pdf, consulté le 05/07/2008

Van Rijsbergen, R.J. (1979), *Information Retrieval*. Butterworths, London (UK), 1979, disponible en ligne : http://www.dcs.gla.ac.uk/Keith/Preface.html, consulté le 05/07/2008

Vickery, B. (1957) Summury of the discussions, Proceedings of the *international conference on classification for information retrieval*, Dorking England, Mai 1957, P.91-108

Werlich, E. (1976), *A text grammar of English*, Heidelberg: Quelle and Meyer

Whissell, C. M. (1989). The dictionary of affect and language. In Plutchik, R. and Kellerman, H., editors, Emotion: Theory, Research, and Experience. volume 4: The measurement of emotions, pages 113-131. Academic Press, New-York.

Wiebe, J. (2000). "Learning Subjective Adjectives from Corpora", In Proceedings of the *AAAI-00*, disponible en ligne : http://www.cs.pitt.edu/~wiebe/pubs/papers/aaai2000.pdf, consulté le 05/07/2008

Wilson, P. (1973), Situational relevance. Information Storage & Retrieval, *American Society for Information Science*, 26(6), pp.321-343.).

Xiaochuan, N., Gui-Rong, X., Xiao, L., Yong, Y., Qiang, Y. (2007) "exploring in the weblog space by detecting informative and affective articles",

International World Wide Web Conference 2007, Canada, p. 281-290, disponible en ligne : http://www2007.org/papers/paper225.pdf, consulté le 05/07/2008

Yi, J., T. Nasukawa, R. Bunescu, Niblack, W. (2003). "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques". *In IEEE ICDM*, disponible en ligne : http://www.cs.utexas.edu/~razvan/papers/icdm2003.pdf, consulté le 05/07/2008

Zhicheng, D., Ruihua, S., Ji-Rong, W. (2007), "A large-scale evaluation and analysis of personalized search strategies", *Proceedings of the 16th international conference on World Wide Web*, May 08-12, 2007, Banff, Alberta, Canada, disponible en ligne : http://www2007.org/papers/paper495.pdf, consulté le 05/07/2008

B. Publications de l'auteur depuis 1994

1. QUONIAM L., ROSTAING H., BOUTIN E., DOU H, 1995, Treating bibliometric indicators with caution : their dependance on the source database , Research Evaluation, Volume 5, number 3, p.177-181, 1995
2. BOUTIN E, QUONIAM L.,ROSTAING H., DOU H., 1995a, a new approach to display real co-authorship and co-topicship through network mapping, Poster, 5ième conference ISSI, 7-10 juin 1995, Chicago, USA.
3. BOUTIN E., DUMAS P., QUONIAM L., ROSTAING H.,DOU H, 1995b,. Génération automatique de réseaux en bibliométrie, communication aux journées d'études organisées par la SFBA sur les Systèmes d'information élaborée à île Rousse, 31 Mai- 2 Juin1995
4. BOUTIN E., QUONIAM L., ROSTAING H., DUMAS P, 1996b. Traitement de l'information : analyse de données classiques versus analyse de réseau. Un cas d'application : la bibliométrie , Actes du dixième congrès national des Sciences de l'information et de la communication, Information , communication et technique : regard sur la diversité des enjeux, Grenoble, 1996, p.571-587
5. BOUTIN E., DUMAS P., ROSTAING H., QUONIAM L, 1996a, . Les réseaux comme outils d'analyse en bibliométrie. Un cas d'application : les réseaux d'auteurs les Cahiers de la documentation Belge, 1996, N°1, p.3-13
6. BOUTIN E., ROSTAING H., QUONIAM L., 1997a, Audit d'un serveur Internet et approche réseau , Colloque de la société française de bibliométrie appliquée : Les systèmes d'information élaborés, île Rousse, 12-16 Mai 1997.
7. BOUTIN E., FERRANDI J.M., VALETTE FLORENCE P., 1997b, Means-end chain Model and Automatic Network Drawing as Commercial Awareness Tools , International Journal of Information Sciences for Decision Making, April 1997, p.19-34
8. BOUTIN E., MANNINA B., ROSTAING H., QUONIAM L.,1998, Construction automatique de réseaux : un outil pour mieux appréhender l'information provenant de l'internet 4èmes journées internationales d'analyse Statistiques des données textuelles, JADT 1998, Nice version pdf
9. ROSTAING H., ZIEGELBAUM H., BOUTIN E., ROGEAUX M., QUONIAM L. Analyse de commentaires libres par la technique des réseaux de segments , Actes des 4èmes journées internationales d'analyse Statistiques des données textuelles, JADT 1998, p. 695-704
10. ROSTAING H., BOUTIN E., 1999dEvaluation des sites internet. Construction d'indicateurs à partir des réseaux de sitations , Communication acceptée, Colloque les Systèmes d'information élaborée, ile Rousse 27 Septembre-1 Octobre 1999. version pdf
11. BOUTIN E, FERRANDI JM, 1999a, A measuring instrument of the audience of a web site : the network analysis , The 15th IMP Conference, Dublin, 2-4 Septembre 1999.
12. BOUTIN E, FERRANDI JM, 1999b, L'analyse réseau : un outil de mesure de l'audience d'un site internet , colloque la recherche en Iut, Aix en provence, 1999
13. BOUTIN E, OPRIME P. , RENUCCI F., GASTE D., 1999c, La cartographie de sites Web : un outil pour mieux appréhender l'information provenant de l'internet , Communication acceptée, Colloque Hypertexte et hypermédias, Paris 23-24 Septembre 1999 version pdf
14. BERTACCHINI Y., DUMAS P., BOUTIN E., 2000, Vers une représentation de l'état des liens des acteurs locaux , 5e Journées Internationales d'Analyse des Données Textuelles, Ecole Polytechnique Fédérale, Lausanne, mars 2000
15. BOUTIN E., FERRANDI Jean Marc, VALETTE FLORENCE Pierre, 2000, La mesure a priori de l'influence D'une modification d'un site web Sur la répartition de l'audience Entre ses pages : Modélisation et application Au site intranet du fret de la sncf , Communication acceptée, Colloque AFM 2000, Montréal 19 Mai 2000 version pdf

16. FERRANDI JM, BOUTIN E. Application de l'analyse réseau à la modélisation de la visite d'un site web, Recherche et Applications en Marketing, Vol.16, n° 3/2001
17. DUMAS P., BULINGE F., BOUTIN E., "Ethical dimensions of KM in professionnal settings" [KM Brésil 2002]
18. BOUTIN E., Méthodologie relationnelle d'extraction de connaissances à partir de données provenant d'un forum de discussion , Revue : Information, Savoirs, Decisions et Mediations (ISDM) N° 9, 2003
19. PANTEADO Roberto, DOU Henri, BOUTIN E., QUONIAM Luc, 2003b"De la création des bases de données au développement de systèmes d'intelligence pour l'entreprise , Revue : Information, Savoirs, Decisions et Mediations (ISDM), N° 8, 2003
20. PANTEADO R., DOU H., BOUTIN E., QUONIAM L., 2003a"Da criacao de bases de dados ao desenvolvimento de sistemas de inteligência para a organizacao", 4eme Workshop Brésilien d'Intelligence Compétitive et Gestion de la Connaissance, 20, 21, 22 octobre 2003, San salvador – Brésil, 2003
21. GASTE D., BIRIOUKOFF E., BOUTIN E., 2003b La bourse de compétences : un dispositif combinant médiation des savoirs et médiatisation des compétences, Colloque : TICEMED, Revue : Information, Savoirs, Decisions et Mediations (ISDM), 2003
22. BERTACCHINI Y., BOUTIN E., Une lecture possible du territoire Sophipolitain : l'observation des représentations virtuelles d'une technopole , in Information, Savoirs, Decisions et Mediations (ISDM), 7, 2003
23. BOUTIN E., 2004a L'exploitation de l'âge d'une page web : quelles perspectives pour l'analyse cybermétrique , colloque, VSST 2004, tome B, P.439-449
24. BOUTIN E., 2004b Qualifier la présence d'une ville sur le web par des indicateurs cybermétriques dynamiques : une expérimentation sur 10 villes françaises , colloque TIC et territoire, quels développements ?, Lille, 2004
25. BOUTIN E., GASTE D., 2004, "Exploitation d'une bourse de compétences pour la construction d'un bilan de compétences , Colloque : TICEMED, Revue : Information, Savoirs, Decisions et Mediations (ISDM), 2004
26. PIERRET J D, BOUTIN E., 2004 Découverte de connaissances dans les bases de données bibliographiques. Le travail de Don Swanson : de l'idée au modèle , Revue : Information, Savoirs, Decisions et Mediations (ISDM), N°12, 2004
27. PIERRET J. D., DOLFI F., QUONIAM L., BOUTIN E., RICCIO E., 2005 Découverte de connaissances dans les bases de données bibliographiques. Modèles expérimentaux autour de la première hypothèse de Swanson , Revue : Information, Savoirs, Decisions et Mediations (ISDM), Isdm N°20, 2005
28. BOUTIN E., MARTAILLAN S., caractérisation du maillage territorial sur Internet : une validation expérimentale en région PACA , 5ième assises de l'internet, Nice, 2005
29. BOUTIN E., CADEL P., 2005, Qualifier la présence d'une ville sur le Web par des indicateurs cybermetriques spatio-temporels : une validation expérimentale pour 2 villes moyennes de la région de Tunis , Colloque : Institut Supérieur de la Documentation, Colloque international, Tunis, 2005
30. BOUTIN E., PERRIN G., 2005, Construction du réseau d'interaction entre sites web : test de robustesse de la méthode à partir de plusieurs sources d'information , colloque ile rousse, 2005
31. DUMAS P., BOUTIN E., DUVERNAY D., GALLEZOT G., 2005, "is communication separable from information , First european communication conference, Amsterdam, 2005
32. PERRIN G., BOUTIN E., 2005, Représentation et analyse des interactions entre les acteurs du web public régional en région PACA : un outil au service de l'intelligence territoriale , colloque ile rousse, 2005
33. PANTEADO R., QUONIAM L., BOUTIN E., DOU H., FARIA L., 2005, "Inteligencia Competitiva na Análise Estrategica das Competencias: o caso Embrapa", sixieme Conferencia Anual de Inteligencia Competitiva. Brésil, Sao Paulo: IBC Brésil, 26-28 avril, Sao Paulo, 2005
34. ROMMA N., BOUTIN E., 2005b, Les stratégies d'influence sur Internet : validation expérimentale sur le lobby antinucléaireColloque : Les systèmes d'information élaborée (Ile Rousse), 2005

35. BOUTIN E., ROMMA N., 2005 le marché des outils de recherche majeurs. Entre stratégies des acteurs et perception des internautes : une communication d'influence virale , 2005 [Le monde selon Google, Googling or not Googling, Bucarest, Juin 2005]
36. BOUTIN E., 2005, Analyse réseau et traitement de données massives : quelles perspectives – Rencontres 2005 du pôle méthodologiqueAnalyse des données relationnelles EHESS, 2005 [Rencontres 2005 du pôle méthodologique Analyse des données relationnelles EHESS-INED]
37. CADEL P., BOUTIN E., 2005, les spécificités du web : un obstacle à son exploitation , 2005 [Institut Supérieur de la Documentation, Colloque international, Tunis]
38. ROMMA R., BOUTIN E., 2006, internet entre homogénéisation et diversité culturelle : comparatif du web français et russe , 2006 [deuxième conférence francophone en science de l'information et de la communication en Russie Médias et diversité culturelle , Moscou]
39. BOUTIN E., Biais cognitifs et recherche d´information sur internet. Quelles perspectives pour les indicateurs de pertinence des moteurs de recherche , Colloque : VSST, 2006
40. BOUTIN E., GALLEZOT G., QUONIAM Luc, 2006a, Détecter l'innovant sur le web par des techniques non booléennes : méthode, outils, application , Colloque : colloque Canadien des sciences de l'information, Disponible en ligne : http://www.cais-acsi.ca/proceedings/2006/boutin_2006.pdf, 2006
41. BOUTIN E., QUONIAM Luc, Thésaurus et clusterisation automatique de données web : deux outils au service de la détection de signaux faibles , Colloque : colloque Contecsi, Brésil, 2006
42. BOUTIN E., QUONIAM Luc, GALLEZOT G., 2006b,Nouvelle approche classificatoire appliquée au web, Une validation expérimentale : représentation des sciences de l'information et de la communication sur le web , Colloque : ISKO - Pratiques et méthodes de classification du savoir à l'heure d'Internet – Nantes, 2006
43. GALLEZOT G., DUMAS Philippe, BOUTIN E., 'Les Sciences de l'Information ET de la Communication : une problématique du et , 15ième congrès SFSIC Bordeaux 10-12 Mai 2006, 2006
44. PINCZON DU SEL P., DUMAS P., BOUTIN E., L'UTILISATION DES TIC EN INTELLIGENCE ECONOMIQUE : LE REVERS DE LA MÉDAILLE , in Degres, 01/06/2006
45. PERRIN G., BOUTIN E., Political administrations on the internet: what kind of territorial network representation", Colloque : Proceedings of the IADIS International Conference: e-Society 2006, Dublin, July 2006. IADIS Press, 2006
46. PERRIN G., BOUTIN E., MARTIN O., 2006, "Couleurs de pages web : à la recherche du lien réel virtuelColloque : VSST Lille, 2006
47. PANTEADO Roberto, BOUTIN E., 2007, "Emerging Technologies of Text Mining: Techniques and Applications", ouvrage collectif, 2007
48. PANTEADO R., VILCHES NOGUEIRA J.H., CORREA DA FONSECA JUNIOR W., CIPRIANO R., QUONIAM L., BOUTIN E., 2007a, Mediametry: Monitoring functions of the organizations communication with bibliometric methodologies and softwares , Dr. Edson Luiz Riccio, SAO PAOLO Brésil, 2007 [colloque Contecsi (Congresso Internacional de Gestão da Tecnologia e Sistemas de Informação)]
49. PANTEADO R., VILCHES NOGUEIRA J.H., CORREA DA FONSECA JUNIOR W., CIPRIANO R., QUONIAM L., BOUTIN E., 2007b Mediametry: Monitoring functions of the organizations communication with bibliometric methodologies and softwares , in Revue Prisma, SAO PAOLO, 2007
50. BERTACCHINI Y., DUMAS P., HERBAUX P., VENTURINI MM, BOIS C., BOUTIN E., Intelligence territoriale, le Territoire dans tous ses états , Toulon : Presses technologiques Coll.les ETIC, 10/07
51. BOUTIN E., LIU L., YUAN Y., 2007a, Les réseaux latents : un outil au service de l'intelligence économique , 2007 [VSST 2007]
52. LIU P, BOUTIN E., DUVERNAY D., DUMAS P., 2008, Les thésaurus comme outils de repérage de la diversité culturelle dans les pratiques d'une discipline. Approche expérimentale dans le champ de l'intelligence économique, 2008 [SFSIC-ISD : Tunis

Les valeurs de l'interaction et de la transmission dans les sciences de l'information et de la communication]

53. BOUTIN E., LIU P., VU M.C., NGUYEN H., HOANG M.D., 2008a, L'Intelligence Economique en Asie et en occident : différences culturelles , paris : CCI, 2008 [Intelligence Economique et Francophonie vecteur de développement et de coopération internationale]
54. ERTZSCHEID O ;, GALLEZOT G., BOUTIN E., 2008, Perspectives documentaires sur les moteurs de recherche : entre sérendipité et logiques marchandes , article dans ouvrage collectif Google…, à Paraître 2008
55. BOUTIN E., GALLEZOT G., Duvernay Daphné, 2008b, Les réseaux latents : adjuvants de notre unité plurielle , communication acceptée, colloque SFSIC 2008
56. BOUTIN E., LIU P., GORIA S., DUMAS P., DAVID A., 2008c, Le domaine de l'information – communication vu à travers le thésaurus Rameau : reflet de la réalité ou interprétation du monde , communication acceptée, colloque SFSIC 2008
57. BOUTIN E., GALLEZOT G., CADEL P., LIU P., PERRIN G., 2008d, Cartographies des flux virtuels entre les acteurs de la technopole de Sophia Antipolis , article dans ouvrage collectif Réseaux d'innovation - Enjeux de la communication au sein d'une technopole, Le cas Sophia Antipolis , a paraître , 2008
58. YOUSSEF M., BOUTIN E., SOUARI W., 2008a, les outils de dialogue entre les cultures : cas des forums de discussion, Communication acceptée, Colloque Ticemed, Sfax, 2008
59. YOUSSEF M., PERRIN G., BOUTIN E., 2008b, Interactions hypertextuelles entre les sites Web culturels maghrébins , Communication acceptée, Colloque Ticemed, Sfax, 2008
60. LIU P., YUAN Y., BOUTIN E. (2008), 社会网络分析法在论文合作网中的应用研究, Application of Social Network Analysis in Information Science, Journal of the China Society for Scientific and Technical information, Mai 2008
61. LIU P., YUAN Y., BOUTIN E. (2008), 中国情报学论文地区合作网研究, The study of region cooperation network of papers in Chinese information science, journal of information, vol(10) 2008 (octobre)

ANNEXES

1. Annexe 1 : Front de recherche sur facet theory

Beghtol C, (1995), «Facets» as interdisciplinary undiscovered public knowledge: S.R. Ranganathan in India and L. Guttman in Israel, Journal of Documentation, 1995, vol. 51, no3, pp. 194-224

Broughton, V. (2006). The need for a faceted methods of information retrieval. ASLIB Proceedings, 58(1-2), 49-72

Coleman, M., Liau, T. L. (1975); A computer readability formula designed for machine scoring, Journal of Applied Psychology, Vol. 60, pp. 283-284 zzaa a trouver

Dale, Chall J.S, 1948, "A Formula for Predicting Readability," Educational Research Bulletin 27 21 Janvier 21 (pp. 11-20) et 17Février (pp.37-54) zzaa à trouver.

Davis I (2005) Why tagging is expensive, Blog disponible à http://blogs.talis.com/panlibus/archives/2005/09/why_tagging_is.php (accessible le 23 MArs 2008)

Devadason Francis J., Intaraksa Neelawat; Patamawongjariya Ponprapa, Desai Kavita, (2002), Faceted indexing based system for organizing and accessing Internet resources, Knowledge organization , 2002, vol. 29, n°2, pp. 65-77, http://www.geocities.com/devadason.geo/FactDSIS.pdf

Ellis, David. "Ranganathan and the Net: using facet analysis to search and organise the World Wide Web." Aslib Proceedings 51 (January 1999): 3-10

Gödert, W. (1991), "Facet classification in online retrieval", International Classification, Vol. 18 No.2, pp.98-109.

Kashyap, M. M., (2003), Likeness Between Ranganathan's Postulations Based Approach to Knowledge Classification and Entity Relationship Data Modelling Approach, Knowledge Organization, Vol. 30 (2003), No. 1

Kim, Jeong-Hyen; Lee, Kyung-Ho, (2002), Designing a Knowledge Base for Automatic Book Classification., Electronic Library, v20 n6 p488-95 2002

La Barre, Kathryn (2004), The art and science of classification: Phyllis Allen Richmond, 1921-1997., Library Trends, 2004

Mills, Jack, (2004), Faceted Classification and Logical Division in Information Retrieval, In Library Trends 52(3) Winter 2004: 541-570, http://www.ideals.uiuc.edu/bitstream/2142/1687/2/Mills541570.pdf

Muhittin Oral and Ossama Kettani, The facets of the modeling and validation process in operations research, European Journal of Operational Research, 1993, vol. 66, issue 2, pages 216-234

Lancaster (F W); Zeter (M J);Metzler (L). Ranganathan's influence examined bibliometrically. Libri, Vol.42 (3); 1992; p 268 – 281

Neelameghan A., (1992), Application of Ranganathan's general theory of knowledge classification in designing specialized databases, Libri, 1992, vol. 42, no3, pp. 202-226

Sharma, R.N., (1992) "Ranganathan's Impact on International Librarianship Through Information Technology. "Libri (International Library Review), 42(3): 258-267 (1992).

Shirky C., 2005, "Ontology is Overrated: Categories, Links, and Tags", http://shirky.com/writings/ontology_overrated.html, consulté le 25 Mars 2008

Spiteri, Louise. "A simplified model for facet analysis: Ranganathan 101". Canadian Journal of Information and Library Science 23 (April-July 1998): 1-30.

STONE A. T, That elusive concept of aboutness': the year's work in subject analysis, 1992, Library resources & technical services , 1993, vol. 37, no3, pp. 277-298

Uddin Mohammad Nasir, Janecek Paul (2007), Faceted classification in web information architecture: A framework for using semantic web tools, The Electronic Library, Vol 25, P219-233

Van Der Walt Marthinus S. (2004), A classification scheme for the organization of electronic documents in small, medium and micro enterprises (SMMES), Knowledge organization, 2004, vol. 31, n°1, pp. 26-38

Zins Chaim, Guttmann David (2000), Structuring Web bibliographic resources : An exemplary subject classification scheme, Knowledge organization , 2000, vol. 27, no3, pp. 143-159.

Zins Chaim; Gutmann David (2003), Domain analysis of social work: An example of an integrated methodological approach , Knowledge organization, 2003, vol. 30, n°3-4, pp. 196-212

2. Annexe 2 : Bases intellectuelles sur facet theory

Austin, D. (1984). PRECIS: A manual of concept analysis and subject indexing (2nd ed.). London: British Library.

Ellis, D and Vasconcelos, A. (1999). Ranganathan and the Net : using facet analysis to search and organize the world wide web. (Aslib Proceedings. . 51 (1). 3- 10).

Ellis David , VASCONCELOS Ana, (2000), The relevance of facet analysis for World Wide Web subject organization and searching, Journal of internet cataloguing, 2000 vol. 2 , no 3-4 , pp. 97 - 114

Foskett, A.C. (1996). The subject approach to information. 5th ed. London: Library Association.

Hjorland, Birger. (1998), Information Retrieval, Text Composition, and Semantics. Knowledge Organization. l99g; 25(L /2): 1G31.' ISSN: 0943- 744.

Ingwersen, P. & Wormell, I. (1992). Ranganathan in the perspective of advanced information retrieval. Libri, 42, 184-201.

Mills, J., & Broughton, V. (1977-). Bliss bibliographic classification (2nd ed.). London: Butterworths.

Neelameghan a, (1991), concept categorization and knowledge organization in specialized databases - a case-study, international classification 18 (2): 92-97

Ranganathan, S. R. (1960). Colon classification: Basic classification (6th ed.). London: Asia Publishing House.

Ranganathan, S. R. (1933). Colon Classification. Madras: Madras Library Association. (1st edition).

Ranganathan, S. R. (1937). Prolegomena to Library Classification. The Madras Library Association. 2nd Ed, The Madras Library Association, 1957. 3rd edition. London: Asia Publishing House.

Ranganathan, Shiyali Ramamrita. "The Five Laws of Library Science", Delhi, Ess Ess Publications (Reprint) 2007.

Satija MP, (1992), Ranganathan and classification - a chronology 1924-1992, international classification 19 (1): 3-6

StilesWG, (1985), Ranganathan, cognition and expert systems, canadian journal of information science-revue canadienne des sciences de l information 10: 16-24

Tversky, A. ; Kahneman, D., 1974, " Judgment under uncertainty: heuristics and biases. ", *Science*, Vol. 185, pp. 1124–1131. ()

Vickery B. C (1958). - Classification and indexing in science. - London, Butterworths Publ., 1958

Vickery, B.C. 1960. Faceted classification: a guide to construction and use of special schemes. London: Aslib.

3. Annexe 3 : facette valence – jeu de calibrage

			très <	plutôt nég	neutre	plutôt >0	très >
Nucléaire	3	http://www.sortirdunucleaire.org/					
	1	http://www.total.com/fr/press/press_releases/pr_2007/071121-					
	2	http://www.polmar.com/pollution/petrole.htm					
	4	http://www.total.com/fr/responsabilite-societale-environnemen					
Total	5	http://www.seasailsurf.com/seasailsurf/actu/spip.php?article21					
	2	http://rumeursdunet.com/nicolas-sarkozy-plus-viril-qu-un-acte					
Sarkozy	4	http://www.syti.net/SarkozyDanger.html					
	1	http://goji.over-blog.net/					
	3	http://www.gojisante.net/					
	4	http://www.aquadesign.be/news/article-11931.php					
Goji	5	http://www.boutique-bonheur.com/548/Authentiques-baies-de-					
	1	http://www.regime-facile.net/					
	2	http://www.servicevie.com/02SANTE/Manchette/Manchette08					
	3	http://www.regime-gratuit.com/					
	4	http://www.cigogne-loutre.com/html/regimeloutre.html					
Régime	5	http://forum.aufeminin.com/forum/sante3/__f1326_sante3-Arr					
	2	http://tempsreel.nouvelobs.com/actualites/international/asiepa					
	3	http://www.cyberpresse.ca/article/20071120/CPMONDE/7112					
dalai Lama	5	http://www.tibetlibre.org/DossiersETL/Le%20Panchen%20Lan					
	3	http://www.tabac-cigarette.com/					
	4	http://fr.news.yahoo.com/afp/20080102/tfr-sante-tabac-restau					
Fumer	5	http://www.lefigaro.fr/debats/2007/12/31/01005-20071231ART					
	3	http://www.afrik.com/article8223.html					
	4	http://www.hemes.be/esas/mapage/euxaussi/famille/polygame					
Polygamie	5	http://www.croixsens.net/mariage/polygamie.php					
	2	http://www.diplomatie.gouv.fr/fr/actions-france_830/crises-cor					
darfour	4	http://www.actioncontrelafaim.org/nos-missions/temoignages/					
	1	http://www.ecofac.org/Canopee/N10/N1003_CITES/CITES_E					
	2	http://www.futura-sciences.com/fr/sinformer/actualites/news/t/					
	3	http://www.miniature.utopia-market.com/index.php?langue=1&					
Elephant	4	http://www.ifaw.org/ifaw/general/default.aspx?oid=213317					
	1	http://nature.ca/discover/treasures/foss/tr1/cha_f.cfm					
	2	http://www.museum-toulon.org/					
	3	http://pythounet.free.fr/dinohist.htm					
	4	http://lyceepasteur-ceb-ccslf.com.br/Pedag/Primaire/litterature					
dinausore	5	http://www.france5.fr/maternelles/loisirs/W00266/35/134334.c					
	1	http://www.doctissimo.fr/html/dossiers/pollution/articles/10120					
	2	http://www.cea.fr/jeunes/themes/le_climat/questions_sur_l_eff					
	3	http://www.manicore.com/documentation/serre/gaz.html					
	4	http://www.planetecologie.org/kitdevdur/effser.html					
Effet de serre	5	http://www.sur-la-toile.com/mod_News_article_133___.html					
	1	http://www.ldi5.com/phys/psyche%20quantique.php					
	2	http://www.paranormal-info.com/Dossier-La-synchronicite-une					
Synchronicité	4	http://en.marge.free.fr/intuit_007.htm					
	2	http://www3.sympatico.ca/gaston.ringuelet/lepetitprince/chapit					
	3	http://www.ac-creteil.fr/lettres/pedagogie/college/6e/petit_prin					
	4	http://www.lepetitprince.com/fr/REVERB/rose_PP.php					
Le petit prince	5	http://papyrusa.travelblog.fr/194727/La-paix-de-la-Rose-du-pe					
	1	http://perso.dixinet.com/animaux-infos/cameleon.html					
	3	http://www.momes.net/dictionnaire/c/cameleon.html					
	4	http://www.fotocommunity.fr/pc/pc/display/9352398					
Caméléon	5	http://www.annoncesetanimaux.com/forum/sujet-6859.html					

Tableau 48 : liste des 52 pages constituant le test de calibrage

4. Annexe 4 : facette valence – résultat calibrage

num	page	très négative	plutôt négative	neutre	plutôt positive	très positive	total	evaluation
1	http://www.seasailsurf.com/seasailsurf/actu/spip.php?article2114	2	4	1	0	1	8	négatif
2	http://www.syti.net/SarkozyDanger.html	14	0	1	0	0	15	négatif
3	http://rumeursdunet.com/nicolas-sarkozy-plus-viril-qu-un-acteur-porno	4	10	9	1	1	25	négatif
4	http://www.tabac-cigarette.com/	24	8	1	2	0	35	négatif
5	http://www.lefigaro.fr/debats/2007/12/31/01005-20071231ARTFIG00164-la-cigarette-du-condamne.php	0	1	4	1	0	6	neutre
6	http://perso.dixinet.com/animaux-infos/cameleon.html	0	2	4	1	0	7	neutre
7	http://www.paranormal-info.com/Dossier-La-synchronicite-une.html	0	1	4	3	0	8	neutre
8	http://www.planetecologie.org/kitdevdur/effser.html	0	1	3	4	0	8	neutre
9	http://www.annoncesetanimaux.com/forum/sujet-6859.html	0	3	4	2	0	9	neutre
10	http://www.cyberpresse.ca/article/20071120/CPMONDE/71120054/1033/CPMONDE	0	2	5	1	1	9	neutre
11	http://www.diplomatie.gouv.fr/fr/actions-france_830/crises-conflits_1050/darfour_1091/index.html	1	1	6	0	1	9	neutre
12	http://www.museum-toulon.org/	0	2	3	3	1	9	neutre
13	http://nature.ca/discover/treasures/foss/tr1/cha_f.cfm	0	0	5	4	1	10	neutre
14	http://www.croixsens.net/mariage/polygamie.php	0	3	5	3	1	12	neutre
15	http://www.manicore.com/documentation/serre/gaz.html	0	2	10	1	0	13	neutre
16	http://www.afrik.com/article8223.html	3	4	6	1	1	15	neutre
17	http://www.regime-gratuit.com/	0	6	3	6	2	17	neutre
18	http://www.sortirdunucleaire.org/	1	5	4	7	0	17	neutre
19	http://www.hemes.be/esas/mapage/euxaussi/famille/polygame.html	2	2	12	4	2	22	neutre
20	http://www.ecofac.org/Canopee/N10/N1003_CITES/CITES_ElephantIvoire.htm	1	5	4	2	0	12	neutre
21	http://www.fotocommunity.fr/pc/pc/display/9352398	0	0	1	3	2	6	positif
22	http://papyrusa.travelblog.fr/194727/La-paix-de-la-Rose-du-petit-Prince-le-cinquieme-element/	0	1	1	4	1	7	positif
23	http://www.boutique-bonheur.com/548/Authentiques-baies-de-Goji--1kg-61.00%80.html	0	0	2	4	1	7	positif
24	http://www3.sympatico.ca/gaston.ringuelet/lepetitprince/chapitre21.html	0	0	2	2	4	8	positif
25	http://www.lepetitprince.com/fr/REVERB/rose_PP.php	0	0	0	4	5	9	positif
26	http://www.aquadesign.be/news/article-11931.php	0	1	2	5	3	11	positif
27	http://www.futura-sciences.com/fr/sinformer/actualites/news/t/vie-1/d/ivoire-le-trafic-continue-avec-175-elephants-abattus_9282/	1	2	1	7	0	11	positif
28	http://pythounet.free.fr/dinohist.htm	0	1	1	4	6	12	positif
29	http://www.momes.net/dictionnaire/c/cameleon.html	0	0	3	7	2	12	positif
30	http://www.total.com/fr/press/press_releases/pr_2007/071121-energie-solaire-photovoltech_14206.htm	0	1	4	7	2	14	positif
31	http://www.gojisante.net/	0	0	3	2	9	14	positif
32	http://forum.aufeminin.com/forum/sante3/__f1326_sante3-Arretez-les-regimes.html	4	1	1	1	0	7	non consensuelles
33	http://www.doctissimo.fr/html/dossiers/pollution/articles/10120-effet-de-serre.htm	1	2	2	0	4	9	non consensuelles
34	http://www.total.com/fr/responsabilite-societale-environnementale/environnement-1	0	2	1	3	3	9	non consensuelles
35	http://www.miniature.utopia-market.com/index.php?langue=1&ref_type=F&n_matiere=12&option=99%20	1	4	2	3	0	10	non consensuelles
36	http://www.servicevie.com/02SANTE/Manchette/Manchette08092003/manchette08092003.html	3	2	0	5	0	10	non consensuelles
37	http://www.polmar.com/pollution/petrole.htm	4	4	2	1	1	12	non consensuelles
38	http://www.regime-facile.net/	1	5	1	6	2	15	non consensuelles

5. Annexe 5 : facette valence – jeu de test

	requête	urlcomplète	très nég	plutôt nég	neutre	plutôt pos	très pos		très subje	plutôt subj	plutôt obj	très objective
1		http://www.camptocamp.org/forums/viewtopic.php?pid=8										
2	5 fruits et légur	blagues-en-blog.blogspot.com\2008\01\5-fruits-5-legume										
3	5 fruits et légur	forum.elle.fr\htm2\vos-astuces-pour-consommer-5-fruits-										
4	5 fruits et légur	pofpom.skynetblogs.be/post/3360338/programme-du-mo										
5	5 fruits et légur	www.amazon.fr/méthode-fruits-légumes-par-jour/dp/2702										
6	5 fruits et légur	www.confidentielles.com/ttopic-39512-est-ce-que-vous-n										
7	boycott	culturepolitiquearabe.blogspot.com\2006_09_01_archive										
8	boycott	culturepolitiquearabe.blogspot.com\2007\12\my-life-arab										
9	boycott	www.amazon.fr\boycott-M-Drillech\dp\2878454219										
10	boycott	www.collectifdu29mai.org/No-smoking-presse-Guilvinec-										
11	boycott	www.ism-france.org/news/article.php?id=3769&type=can										
12	boycott	www.wsws.org\francais\News\2002\avril02\26avril02_boy										
13	carla bruni	60millions.viabloga.com/										
14	carla bruni	aixtal.blogspot.com\2008\01\actu-carla-bruni-enceinte.ht										
15	carla bruni	aixtal.blogspot.com\2008\01\rfrencement-carla-bruni-enc										
16	carla bruni	bravepatrie.com\1085-Carla-Bruni-La-modele-etait-une-t										
17	carla bruni	madame.lefigaro.fr\celebrites\en-kiosque\234-carla-bruni										
18	carla bruni	profile.myspace.com/index.cfm?fuseaction=user.viewpro										
19	carla bruni	tf1.lci.fr\infos\france\politique\0,,3652889,00-carla-bruni-										
20	carla bruni	www.amazon.fr/No-Promises-Carla-Bruni/dp/B000L21D\										
21	carla bruni	www.bakchich.info\article2243.html										
22	carla bruni	www.monsieur-biographie.com\celebrite\biographie\carla										
23	carla bruni	www.rue89.com/mon-oeil/nicolas-sarkozy-et-carla-bruni-										
24	carla bruni	www.tdg.ch/pages/home/tribune_de_geneve/l_actu/mon										
25	crise boursiere	www.blogfinel.com\Crise-boursiere-en-ete,-TVA-a-la-rent										
26	crise boursiere	www.boursorama.com/forum/message.phtml?file=36886										
27	crise boursiere	www.britannica.fr\alpha\C\C95.html										
28	crise boursiere	www.cfecgc.org\transfert\Documents_255CBombeFinan										
29	crise boursiere	www.edubourse.com/guide/guide.php?fiche=krach-1929										
30	crise boursiere	www.europe2020.org/spip.php?article421										
31	crise boursiere	www.leap2020.eu/index.php?action=recherche&tag=com										
32	crise boursiere	www.lesafriques.com/international/la-crise-des-subprime										
33	facebook	adscriptum.blogspot.com\2007\10\facebook.html										
34	facebook	fr.wikipedia.org\wiki\Facebook										
35	facebook	mashable.com\2006\08\25\facebook-profile\default.htm										
36	facebook	www.ed-productions.com/leszed/index.php?86-groupes-p										
37	facebook	www.fluctuat.net/6006-Facebook-vs-Meetic										
38	facebook	www.fredcavazza.net\2007\11\14\pourquoi-je-ne-crois-pl										
39	facebook	www.presse-citron.net/?2008/01/05/2912-facebook-petite										
40	ogm	blog.greenpeace.fr\ogm\default.htm										
41	ogm	blog.greenpeace.fr\ogm\projet-de-loi-ogm-mobilisation-g										
42	ogm	http://fr.wikipedia.org/wiki/Organisme_g%C3%A9n%C3%										
43	ogm	lipietz.net/spip.php?mot26										
44	ogm	www.monde-solidaire.org/spip/rubrique.php3?id_rubrique										
45	ogm	www.moratoireogm.fr/spip.php?article1										
46	ogm	www.science-decision.net/cgi-bin/topic.php?topic=ALI										
47	ovni	fr.wikipedia.org/wiki/Liste_des_principales_observations_										
48	ovni	http://fr.wikipedia.org/wiki/Objet_volant_non_identifi%C3										
49	ovni	www.aquadesign.be\news\article-5111.php										
50	ovni	www.cieletespaceradio.fr/index.php/2007/03/22/90-le-do										
51	ovni	www.unice.fr\zetetique\articles\ovni_pilote.html										
52	rapport attali	fr.wasalive.com\fr\rapport+attali										
53	rapport attali	www.betapolitique.fr\Retour-sur-le-rapport-Attali-01825.h										
54	rapport attali	www.int-evry.fr/aphec/article.php3?id_article=232										
55	rapport attali	www.village-justice.com/forum/viewtopic.php?p=138279										
56	voiture électriq	ecologie.caradisiac.com\France-a-quand-la-voiture-elect										
57	voiture électriq	forum.hardware.fr\hfr\Discussions\Auto-Moto\comment-p										
58	voiture électriq	http://fr.wikipedia.org/wiki/V%C3%A9hicule_%C3%A9lec										
59	voiture électriq	http://fr.wikipedia.org/wiki/Voiture_%C3%A9lectrique										
60	voiture électriq	www.20minutes.fr\article\202149\Monde-La-voiture-elect										
61	voiture électriq	www.infosdelaplanete.org\2867\l-assassinat-de-la-voiture										
62	voiture électriq	www.pile-au-methanol.com\roulez-electrique.htm										
63	voiture électriq	www.wipo.int\wipo_magazine\fr\2007\01\article_0001.htn										

Tableau 49 : liste des 63 pages utilisées

6. Annexe 6 : facette valence : résultats du test

	requête	urlcomplète	valence	pagesans traitables sous textpipe pro	très nég	plutôt nég	neutre	plutôt pos	très pos	très subje	plutôt subj	plutôt obj	très objective	valencesynthese	objectivite
1		http://www.camptoca	47	0	4	1				4	1			neg	sub
2	5 fruits et légumes	blagues-en-blog.blog	53	0		1	4	4	3		5	6	1	pos	non consensuelle
3	5 fruits et légumes	forum.elle.fr\htm2\vo	52	0		2	2	5	1	2	5	3		pos	sub
4	5 fruits et légumes	pofpom.skynetblogs.l	45	1		4	3	2		4	2	2		non consensuelle	sub
5	5 fruits et légumes	www.amazon.fr/méth	49	1		5	1	2	1		1	5	3	non consensuelle	obj
6	5 fruits et légumes	www.confidentielles.c	52	1		1	5	6		5	3	4		pos	non consensuelle
7	boycott	culturepolitiquearabe	46	0		2	5	2		1	2	5	1	neutre	non consensuelle
8	boycott	culturepolitiquearabe	47	0	1	2	3	1		1	1	3	2	non consensuelle	obj
9	boycott	www.amazon.fr\boyc	49	1	2	3	3	1		2	1	4	1	neg	non consensuelle
10	boycott	www.collectifdu29ma	44	0	3	3	1	1		4	1	3		neg	non consensuelle
11	boycott	www.ism-france.org/r	47	1	4	3		1		5	2	1		neg	sub
12	boycott	www.wsws.org\franca	42	1	3	2		1	1	3	4			non consensuelle	sub
13	carla bruni	60millions.viabloga.c	44	1	7	3		1		8	3			neg	sub
14	carla bruni	aixtal.blogspot.com\2	46	0		4	3			1	2	4		neg	non consensuelle
15	carla bruni	aixtal.blogspot.com\2	46	0	1	2	3			2	3	1		neg	sub
16	carla bruni	bravepatrie.com\108	45	1	1	3		1		1	2	2		neg	non consensuelle
17	carla bruni	madame.lefigaro.fr\c	52	0			3	2	1	3	1	3		pos	non consensuelle
18	carla bruni	profile.myspace.com	50	0			1	3	1	1	3	1		pos	sub
19	carla bruni	tf1.lci.fr\infos\france\	52	0		1	1	1			2	1		nbinsuff	nbinsuff
20	carla bruni	www.amazon.fr/No-P	50	1				4	1		1	2	2	pos	obj
21	carla bruni	www.bakchich.info\ar	43	0	2	1				3	1			nbinsuff	nbinsuff
22	carla bruni	www.monsieur-biogra	50	1			1	2			1	2		nbinsuff	nbinsuff
23	carla bruni	www.rue89.com/mon	50	0	1	4				5				neg	sub
24	carla bruni	www.tdg.ch/pages/ho	48	1	1	1				2				nbinsuff	nbinsuff
25	crise boursiere	www.blogfinel.com\C	48	1					2			1	1	nbinsuff	nbinsuff
26	crise boursiere	www.boursorama.com	46	1			3	2				3	2	neutre+	obj
27	crise boursiere	www.britannica.fr\alp	37	1	1		1	1					3	nbinsuff	nbinsuff
28	crise boursiere	www.cfecgc.org\trans	44	1										nbinsuff	nbinsuff
29	crise boursiere	www.edubourse.com	47	1		2	1	1	1	1		4		non consensuelle	obj
30	crise boursiere	www.europe2020.org	43	0	1	1	1					2	1	nbinsuff	nbinsuff
31	crise boursiere	www.leap2020.eu/inc	43	1				1				1		nbinsuff	nbinsuff
32	crise boursiere	www.lesafriques.com	38	1	1	2	1	2	1		2	2	3	non consensuelle	obj
33	facebook	adscriptum.blogspot.r	51	0			4	6			5	3	1	pos	non consensuelle
34	facebook	fr.wikipedia.org\wiki\F	50	0		2	3		2			4	3	non consensuelle	obj
35	facebook	mashable.com\2006\	49	0	1		2			1		2		nbinsuff	nbinsuff
36	facebook	www.ed-productions.r	48	1		2	3	1		2	3	1		neutre	sub
37	facebook	www.fluctuat.net/600	51	1		4	3	2		1	3	4	1	non consensuelle	non consensuelle
38	facebook	www.fredcavazza.net	48	0	1				1	1				nbinsuff	nbinsuff
39	facebook	www.presse-citron.ne	51	0	2	1	2	1	1	6	1			non consensuelle	sub
40	ogm	blog.greenpeace.fr\o	46	0	1					1				nbinsuff	nbinsuff
41	ogm	blog.greenpeace.fr\o	45	0			3			1		2		nbinsuff	nbinsuff
42	ogm	http://fr.wikipedia.org	46	0		1	6	1	1		2	3	4	neutre	obj
43	ogm	lipietz.net/spip.php?r	44	0		1	1			1		1		nbinsuff	nbinsuff
44	ogm	www.monde-solidaire	46	1		1				1				nbinsuff	nbinsuff
45	ogm	www.moratoireogm.fr	46	0	1			1		1		1		nbinsuff	nbinsuff
46	ogm	www.science-decisio	48	1		1	1	1			1	1	1	nbinsuff	nbinsuff
47	ovni	fr.wikipedia.org/wiki/L	47	0			5	3	1			3	6	pos	obj
48	ovni	http://fr.wikipedia.org	47	0			2	1			1	1	1	nbinsuff	nbinsuff
49	ovni	www.aquadesign.be\v	47	1	1	1	4		1		2	2	2	neutre	obj
50	ovni	www.cieletespacerad	49	0		2	2	1			3	1	1	neutre	non consensuelle
51	ovni	www.unice.fr\zetetiqu	49	1		2	3				5			neutre-	sub
52	rapport attali	fr.wasalive.com\fr\ra	47	0	1		2	1		1	1	2		nbinsuff	nbinsuff
53	rapport attali	www.betapolitique.fr\	41	0	1		2	1		3	1			nbinsuff	nbinsuff
54	rapport attali	www.int-evry.fr/aphe	49	1			1					1		nbinsuff	nbinsuff
55	rapport attali	www.village-justice.c	50	1	1	2		2	1	3	2		1	non consensuelle	sub
56	voiture électrique	ecologie.caradisiac.c	47	1			1	5	1			5	1	pos	obj
57	voiture électrique	forum.hardware.fr\hfr	46	0		1	3	1		2	2	1		neutre	sub
58	voiture électrique	http://fr.wikipedia.org	45	0			3					2	1	nbinsuff	nbinsuff
59	voiture électrique	http://fr.wikipedia.org	48	0			3	2				5	1	neutre+	obj
60	voiture électrique	www.20minutes.fr\art	48	1		3	1	1			2	3		non consensuelle	non consensuelle
61	voiture électrique	www.infosdelaplanete	45	1	1	3	3	3	1		2	6	1	non consensuelle	obj
62	voiture électrique	www.pile-au-methan	47	1			4	4	1	5	1	4	1	neutre+	non consensuelle
63	voiture électrique	www.wipo.int\wipo_m	48	0		1	3	6	3	1		6	6	pos	obj

7. Annexe 7 : facette subjectivité – jeu de calibrage

requête	Page	très subje	plutôt subj	plutôt obj	très objective	objectivite	nbredemotsnonamt	nbre adj	nbre de mots au co	nombre de je	nbre de première p	nbre de mots à vale	valence<<0	nous nos notre not	% adj	% mots au conditio	% je	% première person	% mots à valencee	% valence<<0	% nous nos notre n
ogm	blog.greenpeace.fr\ogm\def	1				nbinsuff															
ogm	blog.greenpeace.fr\ogm\pro	1		2		nbinsuff															
rapport attali	fr.wasalive.com\fr\rapport+a	1	1	2		nbinsuff															
ovni	fr.wikipedia.org\wiki\Objet_vola	1	1	1		nbinsuff															
voiture électrique	fr.wikipedia.org\wiki\V_25C3_25A9h	2	1			nbinsuff															
ogm	lipietz.net\spip.php@mot26	1		1		nbinsuff															
facebook	mashable.com\2006\08\25\	1		2		nbinsuff															
carla bruni	tf1.lci.fr\infos\france\politique\0	2	1			nbinsuff															
carla bruni	www.bakchich.info\article22	3	1			nbinsuff															
rapport attali	www.betapolitique.fr\Retour	3	1			nbinsuff															
crise boursiere	www.blogfinel.com\Crise-boursiere-	1	1			nbinsuff	530	46	1	2	3	29	16	10	8,68	0,19	0,38	0,57	5,47	3,02	1,89
crise boursiere	www.britannica.fr\alpha\C\C95.html		3			nbinsuff	905	74	0	0	0	90	88	0	8,18	0,00	0,00	0,00	9,94	9,72	0,00
crise boursiere	www.cfecgc.org\transfert\Documents_255C					nbinsuff	1323	79	12	2	2	67	61	5	5,97	0,91	0,15	0,15	5,06	4,61	0,38
crise boursiere	www.europe2020.org\spip.php@arti	2	1			nbinsuff															
facebook	www.fredcavazza.net\2007\	1				nbinsuff															
rapport attali	www.int-evry.fr\aphec\article.php3@	1				nbinsuff	12046	1264	54	0	0	218	155	2	10,49	0,45	0,00	0,00	1,81	1,29	0,02
crise boursiere	www.leap2020.eu\Crise-systemique	1				nbinsuff	983	97	1	0	0	60	53	1	9,87	0,10	0,00	0,00	6,10	5,39	0,10
ogm	www.monde-solidaire.org\s	1				nbinsuff	3560	211	2	0	4	93	66	11	5,93	0,06	0,00	0,11	2,61	1,85	0,31
carla bruni	www.monsieur-biographie.com\	1	2			nbinsuff	766	50	3	5	5	16	8	1	6,53	0,39	0,65	0,65	2,09	1,04	0,13
ogm	www.moratoireogm.fr\spip.p	1		1		nbinsuff															
ogm	www.science-decision.net\cgi-b	1	1	1		nbinsuff	1122	83	1	0	0	23	16	0	7,40	0,09	0,00	0,00	2,05	1,43	0,00
carla bruni	www.tdg.ch\pages\home\tri	2				nbinsuff	970	57	4	1	2	32	22	8	5,88	0,41	0,10	0,21	3,30	2,27	0,82
facebook	adscriptum.blogspot.com\2007\	5	3	1		nsdp	998	60	3	18	27	21	13	9	6,01	0,30	1,80	2,71	2,10	1,30	0,90
carla bruni	aixtal.blogspot.com\2008\01	1	2	4		nsdp	930	35	20	18	25	28	24	4	3,76	2,15	1,94	2,69	3,01	2,58	0,43
5 fruits et légumes	blagues-en-blog.blogspot.com\2	5	6	1		nsdp															
carla bruni	bravepatrie.com\1085-Carla	1	2	2		nsdp	1140	79	15	17	22	60	46	27	6,93	1,32	1,49	1,93	5,26	4,04	2,37
boycott	culturepolitiquearabe.blogsp	1	2	5	1	nsdp	454	31	4	3	4	15	11	2	6,83	0,88	0,66	0,88	3,30	2,42	0,44
carla bruni	madame.lefigaro.fr\celebrit	3	1	3		nsdp	172	5	1	17	22	13	7	0	2,91	0,58	9,88	12,79	7,56	4,07	0,00
voiture électrique	www.20minutes.fr\article\20214	2	3			nsdp	2162	133	6	13	28	73	50	35	6,15	0,28	0,60	1,30	3,38	2,31	1,62
boycott	www.amazon.fr\boycott-M-E	2	1	4	1	nsdp	968	43	6	2	4	15	12	14	4,44	0,62	0,21	0,41	1,55	1,24	1,45
ovni	www.cieletespaceradio.fr\index.	3	1	1		nsdp	642	30	2	12	13	17	14	37	4,67	0,31	1,87	2,02	2,65	2,18	5,76
boycott	www.collectifdu29mai.org\N	4	1	3		nsdp	1981	144	5	6	8	114	76	4	7,27	0,25	0,30	0,40	5,75	3,84	0,20
5 fruits et légumes	www.confidentielles.com\tto	5	3	4		nsdp	758	47	4	25	40	34	13	14	6,20	0,53	3,30	5,28	4,49	1,72	1,85
facebook	www.fluctuat.net\6006-Face	1	3	4	1	nsdp	547	30	3	0	1	23	6	1	5,48	0,55	0,00	0,18	4,20	1,10	0,18
voiture électrique	www.pile-au-methanol.com	5	1	4	1	nsdp	1364	92	11	48	80	42	37	11	6,74	0,81	3,52	5,87	3,08	2,71	0,81
boycott	culturepolitiquearabe.blogsp	1	1	3	2	obj															
voiture électrique	ecologie.caradisiac.com\France-a-g		5	1		obj	1209	106	9	12	15	50	46	8	8,77	0,74	0,99	1,24	4,14	3,80	0,66
ovni	fr.wikipedia.org\wiki\Liste_des_princ		3	6		obj	968	76	3	0	0	27	24	0	7,85	0,31	0,00	0,00	2,79	2,48	0,00
facebook	fr.wikipedia.org\wiki\Facebook		4	3		obj															
ogm	fr.wikipedia.org\wiki\Organisme	2	3	4		obj	4884	417	26	1	10	156	130	3	8,54	0,53	0,02	0,20	3,19	2,66	0,06
voiture électrique	fr.wikipedia.org\wiki\Voiture_25C3	5	1			obj	639	86	8	0	0	25	19	0	13,46	1,25	0,00	0,00	3,91	2,97	0,00
5 fruits et légumes	www.amazon.fr\m_25E9thode-f	1	5	3		obj	701	29	3	0	0	21	10	16	4,14	0,43	0,00	0,00	3,00	1,43	2,28
carla bruni	www.amazon.fr\No-Promises-C	1	2	2		obj	1151	59	4	2	7	20	15	14	5,13	0,35	0,17	0,61	1,74	1,30	1,22
ovni	www.aquadesign.be\news\articl	2	2	2		obj	570	33	6	4	4	16	15	4	5,79	1,05	0,70	0,70	2,81	2,63	0,70
crise boursiere	www.boursorama.com\forum\messa	3	2			obj	610	48	5	0	2	18	16	2	7,87	0,82	0,00	0,33	2,95	2,62	0,33
crise boursiere	www.edubourse.com\guide\	1		4		obj	590	36	0	0	1	10	9	0	6,10	0,00	0,00	0,17	1,69	1,53	0,00
voiture électrique	www.infosdelaplanete.org\2867	2	6	1		obj	549	49	8	3	4	21	18	7	8,93	1,46	0,55	0,73	3,83	3,28	1,28
crise boursiere	www.lesafriques.com\internatio	2	2	3		obj	697	69	4	1	4	34	33	5	9,90	0,57	0,14	0,57	4,88	4,73	0,72
voiture électrique	www.wipo.int\wipo_magazir	1		6	6	obj	1075	81	3	1	1	34	18	0	7,53	0,28	0,09	0,09	3,16	1,67	0,00
carla bruni	60millions.viabloga.com\de	8	3			sub	1561	126	8	57	78	75	55	9	8,07	0,51	3,65	5,00	4,80	3,52	0,58
carla bruni	aixtal.blogspot.com\2008\01	2	3	1		sub	712	39	5	31	38	19	15	4	5,48	0,70	4,35	5,34	2,67	2,11	0,56
5 fruits et légumes	forum.elle.fr\htm2\vos-astu	2	5	3		sub	449	15	1	14	18	7	4	2	3,34	0,22	3,12	4,01	1,56	0,89	0,45
voiture électrique	forum.hardware.fr\hfr\Discu	2	2	1		sub	1001	80	8	24	29	42	33	0	7,99	0,80	2,40	2,90	4,20	3,30	0,00
	www.camptocamp.org\forur	4	1			sub	1017	79	22	23	26	50	30	1	7,77	2,16	2,26	2,56	4,92	2,95	0,10
5 fruits et légumes	pofpom.skynetblogs.be\pos	4	2	2		sub	739	42	2	38	62	72	40	3	5,68	0,27	5,14	8,39	9,74	5,41	0,41
carla bruni	profile.myspace.com\index.	1	3	1		sub	334	8	1	5	7	12	6	1	2,40	0,30	1,50	2,10	3,59	1,80	0,30
facebook	www.ed-productions.com\le	2	3	1		sub	965	50	4	11	23	23	16	5	5,18	0,41	1,14	2,38	2,38	1,66	0,52
boycott	www.ism-france.org\news\a	5	2	1		sub	2898	230	3	1	1	64	52	22	7,94	0,10	0,03	0,03	2,21	1,79	0,76
facebook	www.presse-citron.net\@20	6	1			sub															
carla bruni	www.rue89.com\mon-oeil\n	5				sub	581	25	4	5	10	25	15	11	4,30	0,69	0,86	1,72	4,30	2,58	1,89
ovni	www.unice.fr\zetetique\articles\	5				sub	739	35	4	0	2	17	13	1	4,74	0,54	0,00	0,27	2,30	1,76	0,14
rapport attali	www.village-justice.com\for	3	2		1	sub	770	50	1	6	7	9	9	4	6,49	0,13	0,78	0,91	1,17	1,17	0,52
boycott	www.wsws.org\francais\Nev	3	4			sub	803	67	4	1	1	26	23	1	8,34	0,50	0,12	0,12	3,24	2,86	0,12

Tableau 50 : liste des pages constituant le test de calibrage

8. Annexe 8 : facette subjectivité – pages consensuelles

requête	Page	très subje	plutôt subj	plutôt obj	très objective	objectivite	nbredemotsnonambigusdanslapag	nbre adj	nbre de mots au conditionnel	nombre de je	nbre de première personne du sing	nbre de mots à valenceextremes	valence<<0	nous nos notre notres	% adj	% mots au conditionnel	% je	% première personne du sing	% mots à valenceextremes	% valence<<0	% nous nos notre notres
boycott	culturepolitiquearabe.blogspot.com	1	1	3	2	obj															
voiture électri	ecologie.caradisiac.com\France-a-			5	1	obj	1209	106	9	12	15	50	46	8	8,77	0,74	0,99	1,24	4,14	3,80	0,66
ovni	fr.wikipedia.org\wiki\Liste_des_prin			3	6	obj	968	76	3	0	0	27	24	0	7,85	0,31	0,00	0,00	2,79	2,48	0,00
facebook	fr.wikipedia.org\wiki\Facebook			4	3	obj															
ogm	fr.wikipedia.org\wiki\Organisme_g		2	3	4	obj	4884	417	26	1	10	156	130	3	8,54	0,53	0,02	0,20	3,19	2,66	0,06
voiture électri	fr.wikipedia.org\wiki\Voiture__25C3			5	1	obj	639	86	8	0	0	25	19	0	13,46	1,25	0,00	0,00	3,91	2,97	0,00
5 fruits et légu	www.amazon.fr\m_25E9thode-fruit		1	5	3	obj	701	29	3	0	0	21	10	16	4,14	0,43	0,00	0,00	3,00	1,43	2,28
carla bruni	www.amazon.fr\No-Promises-Carla		1	2	2	obj	1151	59	4	2	7	20	15	14	5,13	0,35	0,17	0,61	1,74	1,30	1,22
ovni	www.aguadesign.be\news\article-5		2	2	2	obj	570	33	6	4	4	16	15	4	5,79	1,05	0,70	0,70	2,81	2,63	0,70
crise boursier	www.boursorama.com\forum\mess			3	2	obj	610	48	5	0	2	18	16	2	7,87	0,82	0,00	0,33	2,95	2,62	0,33
crise boursier	www.edubourse.com\guide\guide.p	1		4		obj	590	36	0	0	1	10	9	0	6,10	0,00	0,00	0,17	1,69	1,53	0,00
voiture électri	www.infosdelaplanete.org\2867\l-a		2	6	1	obj	549	49	8	3	4	21	18	7	8,93	1,46	0,55	0,73	3,83	3,28	1,28
crise boursier	www.lesafriques.com\international\		2	2	3	obj	697	69	4	1	4	34	33	5	9,90	0,57	0,14	0,57	4,88	4,73	0,72
voiture électri	www.wipo.int\wipo_magazine\fr\20	1		6	6	obj	1075	81	3	1	1	34	18	0	7,53	0,28	0,09	0,09	3,16	1,67	0,00
carla bruni	60millions.viabloga.com\default.htr	8	3			sub	1561	126	8	57	78	75	55	9	8,07	0,51	3,65	5,00	4,80	3,52	0,58
carla bruni	aixtal.blogspot.com\2008\01\rfrenc	2	3	1		sub	712	39	5	31	38	19	15	4	5,48	0,70	4,35	5,34	2,67	2,11	0,56
5 fruits et légu	forum.elle.fr\htm2\vos-astuces-pou	2	5	3		sub	449	15	1	14	18	7	4	2	3,34	0,22	3,12	4,01	1,56	0,89	0,45
voiture électri	forum.hardware.fr\hfr\Discussions\	2	2	1		sub	1001	80	8	24	29	42	33	0	7,99	0,80	2,40	2,90	4,20	3,30	0,00
	www.camptocamp.org\forums\view	4	1			sub	1017	79	22	23	26	50	30	1	7,77	2,16	2,26	2,56	4,92	2,95	0,10
5 fruits et légu	pofpom.skynetblogs.be\post\33603	4	2	2		sub	739	42	2	38	62	72	40	3	5,68	0,27	5,14	8,39	9,74	5,41	0,41
carla bruni	profile.myspace.com\index.cfm@fu	1	3	1		sub	334	8	1	5	7	12	6	1	2,40	0,30	1,50	2,10	3,59	1,80	0,30
facebook	www.ed-productions.com\leszed\in	2	3	1		sub	965	50	4	11	23	23	16	5	5,18	0,41	1,14	2,38	2,38	1,66	0,52
boycott	www.ism-france.org\news\article.pl	5	2	1		sub	2898	230	3	1	1	64	52	22	7,94	0,10	0,03	0,03	2,21	1,79	0,76
facebook	www.presse-citron.net\@2008_2F0	6	1			sub															
carla bruni	www.rue89.com\mon-oeil\nicolas-s	5				sub	581	25	4	5	10	25	15	11	4,30	0,69	0,86	1,72	4,30	2,58	1,89
ovni	www.unice.fr\zetetique\articles\ovni		5			sub	739	35	4	0	2	17	13	1	4,74	0,54	0,00	0,27	2,30	1,76	0,14
rapport attali	www.village-justice.com\forum\viev	3	2		1	sub	770	50	1	6	7	9	9	4	6,49	0,13	0,78	0,91	1,17	1,17	0,52
boycott	www.wsws.org\francais\News\2002	3	4			sub	803	67	4	1	1	26	23	1	8,34	0,50	0,12	0,12	3,24	2,86	0,12

Tableau 51 : liste des pages consensuelles constituant le test de calibrage

9. Annexe 9 : facette subjectivité – test

Num	requete	N° Page	url	très subjectif	plutôt subjectif	plutôt objectif	très objectif	consensuelle	OBJECTIVITE JUGES	OBJECTIVITE CALCULE
1	énergie nucléaire	1	http://www.diplomatie.gouv	2	1	7	3	Oui	OBJ	OBJ
2	énergie nucléaire	2	http://agora.qc.ca/mot.nsf/D			10	2	Oui	OBJ	OBJ
3	énergie nucléaire	3	http://www.cea.fr/jeunes/the		1	5	6	Oui	OBJ	OBJ
4	énergie nucléaire	4	http://www.sortirdunucleaire	5	4	2	1	Oui	SUBJ	inf400
5	énergie nucléaire	5	http://www.debat-energie.g	2		5	5	Oui	OBJ	
6	vache folle	6	http://www.vache-folle.com/		1	10	2	Oui	OBJ	inf400
7	vache folle	7	http://www.science-decision		2	5	6	Oui	OBJ	OBJ
8	vache folle	8	http://www.eutraco.com/sa/		9	2	2	Oui	SUBJ	OBJ
9	vache folle	9	http://la-vache-folle.over-blo	12	1			Oui	SUBJ	SUBJ
10	vache folle	10	http://lesverts.fr/mots.php3	2	9	1		Oui	SUBJ	OBJ
11	théorie de l'engagement	11	http://www.menteur.com/ch	2	3		1	Non		
12	théorie de l'engagement	12	http://www.psychologie-soc		2	3	1	Non		
13	théorie de l'engagement	13	http://www.dokimos.ca/co2		2	3	1	Non		
14	théorie de l'engagement	14	http://1libertaire.free.fr/Sou		4	2		Non		
15	théorie de l'engagement	15	http://www.prevensectes.co	1	4	2	1	Non		
16	Dalai Lama	16	http://www.amis-tibet.lu/Por		4	7	3	Oui	OBJ	inf400
17	Dalai Lama	17	http://www.liberation.fr/actu	1	5	6	2	Non		
18	Dalai Lama	18	http://www.tibet-info.net/ww	1	6	2	5	Non		
19	Dalai Lama	19	http://fr.wikipedia.org/wiki/T		2	7	6	Oui	OBJ	OBJ
20	Dalai Lama	20	http://www.info-sectes.org/r	7	5	1	1	Oui	SUBJ	OBJ
21	Malte	21	http://fr.wikipedia.org/wiki/M		1	6	7	Oui	OBJ	OBJ
22	Malte	22	http://users.compaqnet.be/j		5	9		Oui	OBJ	OBJ
23	Malte	23	http://europa.eu/scadplus/le	1	3	6	4	Non		
24	Malte	24	http://www.diplomatie.gouv	1	1	8	4	Oui	OBJ	OBJ
25	Malte	25	http://www.ambafrance-mt.		1	7	6	Oui	OBJ	OBJ
26	Développement durable	26	http://fr.wikipedia.org/wiki/D				4	Oui	OBJ	OBJ
27	Développement durable	27	http://www.ecologie.gouv.fr		2	2	2	Non		
28	Développement durable	28	http://www.developpement	1	1	2	2	Non		
29	Développement durable	29	http://developpement-durab	3	3			Oui	SUBJ	SUBJ
30	Développement durable	30	http://www.conso-ecolo.fr/	1	2	2	1	Non		
31	Le petit prince	31	http://www3.sympatico.ca/g	1	5	3	4	Non		
32	Le petit prince	32	http://fr.wikipedia.org/wiki/P		4	5	5	Non		
33	Le petit prince	33	http://www.radiofrance.fr/re	4	3	4	2	Non		
34	Le petit prince	34	http://prince.frogcircus.org/	3	6	4		Non		
35	Le petit prince	35	http://www.ac-creteil.fr/lettre	2	1	5	5	Oui	OBJ	OBJ
36	Claude François	36	http://fr.wikipedia.org/wiki/C		3	5	6	Oui	OBJ	SUBJ
37	Claude François	37	http://www.ramdam.com/ar	4	10	3	1	Oui	SUBJ	SUBJ
38	Claude François	38	http://www.lefigaro.fr/musiq	4	5	3	1	Oui	SUBJ	SUBJ
39	Claude François	39	http://www.lesclesjunior.com	4	6	2	2	Oui	SUBJ	inf400
40	Claude François	40	http://rumeursdunet.com/cla	9	3	1		Oui	SUBJ	SUBJ
41	Baie d'Halong	41	http://fr.wikipedia.org/wiki/B				4	Oui	OBJ	OBJ
42	Baie d'Halong	42	http://www.discoveryindoch		1	2	1	Non		
43	Baie d'Halong	43	http://www.vietnamtourism.		1	1	2	Non		
44	Baie d'Halong	44	http://www.halong.com/halo	2		1		Non		
45	Baie d'Halong	45	http://arnaud.lethanh.com/h	2	1	1		Non		
46	loup	46	http://www.momes.net/com	2	1	8	6	Oui	OBJ	inf400
47	loup	47	http://www.loupsdugevauda	2	1	9	4	Oui	OBJ	OBJ
48	loup	48	http://www.peupleloup.info/	3	12		1	Oui	SUBJ	SUBJ
49	loup	49	http://www.planete.org/doss		2	3	11	Oui	OBJ	OBJ
50	loup	50	http://www.euroloup.com/en	2	6	5	4	Non		
51	zimbabwe	51	http://www.chateaudebrou.c	4	2		1	Oui	SUBJ	SUBJ
52	zimbabwe	52	http://oui-europe.over-blog.	4	2	1		Oui	SUBJ	SUBJ
53	zimbabwe	53	http://www.boursorama.com			1	6	Oui	OBJ	OBJ
54	zimbabwe	54	http://www.rfi.fr/actufr/article		1	6		Oui	OBJ	OBJ
55	zimbabwe	55	http://www.lefigaro.fr/flash-a		1	2	3	Non		

10. Annexe 10 : lisibilité, calibrage (Kandel et Moles)

requete	numéro page	url	très difficile	plutôt difficile	plutôt facile	très facile	consensuelle LIS	LISIBILITE humaine	LISIBILITE refermee humaine	Formule Kandel et moles	LISIBILITE AUTOMATIQUE Kandel et Mole
énergie nucléaire	1	http://www.diplomatie.gouv.fr/fr/france_829/lab		5	9		Oui	MOY	MOY	55	MOY
énergie nucléaire	2	http://agora.qc.ca/mot.nsf/Dossiers/Energie_nu		6	6		Oui	MOY	MOY	42	DIFF
énergie nucléaire	3	http://www.cea.fr/jeunes/themes/l_energie_nuc	5	5	1		Oui	DIF	DIF	58	MOY
énergie nucléaire	4	http://www.sortirdunucleaire.org/index.php?me			9	3	Oui	PF	MOY	53	MOY
énergie nucléaire	5	http://www.debat-energie.gouv.fr/site/quest_rep	3	3	4	2	Non				
vache folle	6	http://www.vache-folle.com/	2	7	4		Oui	PD	DIFF	45	DIFF
vache folle	7	http://www.science-decision.net/cgi-bin/topic.pl		9	3	1	Oui	PD	DIFF	61	MOY
vache folle	8	http://www.eutraco.com/sa/vache1.html		2	9	2	Oui	PF	MOY	43	DIFF
vache folle	9	http://la-vache-folle.over-blog.com/categorie-10		6	5	2	Oui	MOY	MOY	71	FAC
vache folle	10	http://lesverts.fr/mots.php3?id_groupe=13&id_r	1	4	8	1	Oui	MOY	MOY	25	DIFF
théorie de l'engagement	11	http://www.menteur.com/chronik/000531.html		1	4	1	Non	PF	MOY		
théorie de l'engagement	12	http://www.psychologie-sociale.com/index.php?	1	3	1	1	Non				
théorie de l'engagement	13	http://www.dokimos.ca/co261.htm		4	1	2	Non				
théorie de l'engagement	14	http://1libertaire.free.fr/Soumission03.html		4	2		Non				
théorie de l'engagement	15	http://www.prevensectes.com/manip1.htm			5		Non				
Dalai Lama	16	http://www.amis-tibet.lu/PortraitDLHTML.html		1	10	3	Oui	PF	MOY	59	MOY
Dalai Lama	17	http://www.liberation.fr/actualite/monde/315868	1	4	8	2	Oui	PF	MOY	55	MOY
Dalai Lama	18	http://www.tibet-info.net/www/Discours-de-SS-l		2	10	2	Oui	PF	MOY	51	MOY
Dalai Lama	19	http://fr.wikipedia.org/wiki/Tenzin_Gyatso			6	7	Oui	FAC	FAC	45	DIFF
Dalai Lama	20	http://www.info-sectes.org/religion/dalailama.ht	1	8	2	2	Oui	PF	MOY	57	MOY
Malte	21	http://fr.wikipedia.org/wiki/Malte	1	1	9	3	Oui	PF	MOY	47	DIFF
Malte	22	http://users.compaqnet.be/jpdruine/malte/histoi		6	5	3	Oui	MOY	MOY	53	MOY
Malte	23	http://europa.eu/scadplus/leg/fr/lvb/e15112.htm	1	11	2		Oui	PD	DIFF	44	DIFF
Malte	24	http://www.diplomatie.gouv.fr/fr/pays-zones-ge		5	7	2	Oui	PF	MOY	9	DIFF
Malte	25	http://www.ambafrance-mt.org/article.php3?id	1	2	9	2	Oui	PF	MOY	55	MOY
Développement durable	26	http://fr.wikipedia.org/wiki/D%C3%A9veloppem			2	1	Oui	FAC	FAC	33	DIFF
Développement durable	27	http://www.ecologie.gouv.fr/-Developpement-du		1	5		Oui	PF	MOY		
Développement durable	28	http://www.developpementdurablelejournal.com		3	3		Non				
Développement durable	29	http://developpement-durable.viabloga.com/			4	2	Oui	FAC	FAC	59	MOY
Développement durable	30	http://www.conso-ecolo.fr/		5	1		Oui	PF	MOY	43	DIFF
Le petit prince	31	http://www3.sympatico.ca/gaston.ringuelet/lepe		2	6	5	Oui	PF	MOY	76	FAC
Le petit prince	32	http://fr.wikipedia.org/wiki/Petit_Prince		1	7	5	Oui	PF	MOY	55	MOY
Le petit prince	33	http://www.radiofrance.fr/reportage/dossiers/ste		5	7	1	Oui	MOY	MOY	74	FAC
Le petit prince	34	http://prince.frogcircus.org/		2	10	1	Oui	PF	MOY	69	MOY
Le petit prince	35	http://www.ac-creteil.fr/lettres/pedagogie/colleg		4	4	5	Oui	PF	MOY	69	MOY
Claude François	36	http://fr.wikipedia.org/wiki/Claude_Fran%C3%A			4	10	Oui	TF	FAC	56	MOY
Claude François	37	http://www.ramdam.com/art/f/claudefrancois.ht		1	6	7	Oui	FAC	FAC	55	MOY
Claude François	38	http://www.lefigaro.fr/musique/2008/02/29/0300		3	7	5	Oui	PF	MOY	60	MOY
Claude François	39	http://www.lesclesjunior.com/rubriques/france/r			3	11	Oui	TF	FAC	71	FAC
Claude François	40	http://rumeursdunet.com/claude-francois-un-m		1	7	4	Oui	PF	MOY	60	MOY
Baie d'Halong	41	http://fr.wikipedia.org/wiki/Baie_d'Along			4		Oui	PF	MOY	53	MOY
Baie d'Halong	42	http://www.discoveryindochina.com/fr/croisieres			2	2	Oui	FAC	FAC	56	MOY
Baie d'Halong	43	http://www.vietnamtourism.com/f_pages/herita		3	1		Non				
Baie d'Halong	44	http://www.halong.com/halongcom/f_pages/hal	2	1	1		Non				
Baie d'Halong	45	http://arnaud.lethanh.com/halong.htm			1	3	Oui	FAC	FAC	54	MOY
loup	46	http://www.momes.net/comptines/loup/promen			1	14	Oui	FAC	FAC	64	MOY
loup	47	http://www.loupsdugevaudan.com/		1	13	2	Oui	PF	MOY	76	FAC
loup	48	http://www.peupleloup.info/spip.php?article248		1	12	3	Oui	PF	MOY	53	MOY
loup	49	http://www.planete.org/dossiers/loup/loup_pres		8	7	1	Oui	MOY	MOY	58	MOY
loup	50	http://www.euroloup.com/enfantsloupsenfantss		3	9	4	Oui	PF	MOY	70	MOY
zimbabwe	51	http://www.chateaudebrou.com/zimbabwe/		1	5	2	Oui	PF	MOY	61	MOY
zimbabwe	52	http://oui-europe.over-blog.com/article-484324			5	2	Oui	PF	MOY	65	MOY
zimbabwe	53	http://www.boursorama.com/forum/message.p	2	2	2	1	Non				
zimbabwe	54	http://www.rfi.fr/actufr/articles/099/article_6367		1	6		Oui	PF	MOY	45	DIFF
zimbabwe	55	http://www.lefigaro.fr/flash-actu/2008/03/18/010		1	1	1	Non				

11. Annexe 11 : lisibilité, test de calibrage

requete	numéro page	url	très difficile	plutôt difficile	plutôt facile	très facile	consensuelle LIS	LISIBILITE	FREQMOY	MOYSYLLAB	lisibilité par fréquence de mots	Lisibilité par nombre de syllabes
énergie nucléaire	1	http://www.dip		5	9		Oui	MOY	14,85	2,44	MOY	MOY
énergie nucléaire	2	http://agora.q		6	6		Oui	MOY	15,87	2,76	MOY	DIF
énergie nucléaire	3	http://www.ce	5	5	1		Oui	DIF	15,04	2,42	MOY	MOY
énergie nucléaire	4	http://www.so			9	3	Oui	MOY	16,19	2,82	MOY	DIF
énergie nucléaire	5	http://www.de	3	3	4	2	Non		12,17	2,67	MOY	MOY
vache folle	6	http://www.va	2	7	4		Oui	DIFF	20,36	2,38	FAC	MOY
vache folle	7	http://www.sc		9	3	1	Oui	DIFF	16,03	2,67	MOY	MOY
vache folle	8	http://www.eu		2	9	2	Oui	MOY	11,14	2,67	DIF	MOY
vache folle	9	http://la-vache		6	5	2	Oui	MOY	15,68	2,51	MOY	MOY
vache folle	10	http://lesverts	1	4	8	1	Oui	MOY	12,87	2,79	MOY	DIF
théorie de l'engagement	11	http://www.me		1	4	1	Non	MOY	19,71	2,67	FAC	MOY
théorie de l'engagement	12	http://www.ps	1	3	1	1	Non		17,83	2,78	MOY	DIF
théorie de l'engagement	13	http://www.do		4	1	2	Non		19,59	2,69	FAC	MOY
théorie de l'engagement	14	http://1libertai		4	2		Non		15,45	2,67	MOY	MOY
théorie de l'engagement	15	http://www.pre			5		Non		16,63	2,76	MOY	DIF
Dalai Lama	16	http://www.an		1	10	3	Oui	MOY	16,01	3,00	MOY	DIF
Dalai Lama	17	http://www.lib	1	4	8	2	Oui	MOY	15,16	2,31	MOY	MOY
Dalai Lama	18	http://www.tib		2	10	2	Oui	MOY	19,15	2,70	FAC	MOY
Dalai Lama	19	http://fr.wikipe			6	7	Oui	FAC	13,58	2,56	MOY	MOY
Dalai Lama	20	http://www.inf	1	8	2	2	Oui	MOY	14,88	2,49	MOY	MOY
Malte	21	http://fr.wikipe	1	1	9	3	Oui	MOY	16,25	2,47	MOY	MOY
Malte	22	http://users.cc		6	5	3	Oui	MOY	13,65	2,26	MOY	MOY
Malte	23	http://europa.(	1	11	2		Oui	DIFF	13,17	2,79	MOY	DIF
Malte	24	http://www.dip		5	7	2	Oui	MOY	17,59	2,56	MOY	MOY
Malte	25	http://www.an	1	2	9	2	Oui	MOY	12,37	2,56	MOY	MOY
Développement durable	26	http://fr.wikipe			2	1	Oui	FAC	11,83	2,79	DIF	DIF
Développement durable	27	http://www.ec		1	5		Oui	MOY	16,01	2,70	MOY	MOY
Développement durable	28	http://www.de		3	3		Non		13,81	2,54	MOY	MOY
Développement durable	29	http://develop			4	2	Oui	FAC	13,98	2,52	MOY	MOY
Développement durable	30	http://www.co		5	1		Oui	MOY	14,32	2,49	MOY	MOY
Le petit prince	31	http://www3.s		2	6	5	Oui	MOY	24,44	2,28	FAC	MOY
Le petit prince	32	http://fr.wikipe		1	7	5	Oui	MOY	20,37	2,24	FAC	FAC
Le petit prince	33	http://www.rac		5	7	1	Oui	MOY	20,64	2,37	FAC	MOY
Le petit prince	34	http://prince.fr		2	10	1	Oui	MOY	20,50	2,08	FAC	FAC
Le petit prince	35	http://www.ac		4	4	5	Oui	MOY	17,46	2,39	MOY	MOY
Claude François	36	http://fr.wikipe			4	10	Oui	FAC	17,30	2,22	MOY	FAC
Claude François	37	http://www.rar		1	6	7	Oui	FAC	18,71	2,41	FAC	MOY
Claude François	38	http://www.lef		3	7	5	Oui	MOY	14,73	2,29	MOY	MOY
Claude François	39	http://www.les			3	11	Oui	FAC	16,73	2,22	MOY	FAC
Claude François	40	http://rumeurs		1	7	4	Oui	MOY	23,66	2,25	FAC	MOY
Baie d'Halong	41	http://fr.wikipe			4		Oui	MOY	11,08	2,34	DIF	MOY
Baie d'Halong	42	http://www.dis			2	2	Oui	FAC	18,18	1,94	MOY	FAC
Baie d'Halong	43	http://www.vie		3	1		Non		16,10	1,96	MOY	FAC
Baie d'Halong	44	http://www.ha	2	1	1		Non		16,68	2,14	MOY	FAC
Baie d'Halong	45	http://arnaud.l			1	3	Oui	FAC	16,40	2,07	MOY	FAC
loup	46	http://www.mo			1	14	Oui	FAC	10,62	1,69	DIF	FAC
loup	47	http://www.lou		1	13	2	Oui	MOY	16,86	2,09	MOY	FAC
loup	48	http://www.pe		1	12	3	Oui	MOY	18,17	2,38	MOY	MOY
loup	49	http://www.pla		8	7	1	Oui	MOY	16,81	2,25	MOY	MOY
loup	50	http://www.eu		3	9	4	Oui	MOY	13,33	2,08	MOY	FAC
zimbabwe	51	http://www.ch		1	5	2	Oui	MOY	15,83	2,41	MOY	MOY
zimbabwe	52	http://oui-euro			5	2	Oui	MOY	14,38	2,38	MOY	MOY
zimbabwe	53	http://www.bo	2	2	2	1	Non		14,22	2,13	MOY	FAC
zimbabwe	54	http://www.rfi.		1	6		Oui	MOY	12,69	2,47	MOY	MOY
zimbabwe	55	http://www.lef		1	1	1	Non				DIF	FAC

12. Annexe 12 : liste des mots vides

Liste des mots vides classés par fréquences d'apparition décroissante dans le corpus de film

je, de, vous, à, et, on, des, nous, me, dans, du, te, mon, au, va, ils, toi, faire, oui, veux, sais, cette, peux, vais, alors, très, ou, fais, jamais, aussi, voir, faut, chose, ta, est-ce que, sa, peut, encore, mes, vraiment, temps, toujours, deux, sans, vas, dis, crois, vie, dois, trop, fois, peut-être, monde, sûr, viens, accord, aux, dieu, homme, besoin, oh, femme, vos, chez, aime, père, ses, veut, quelqu'un, tes, voilà, quel, ans, vois, beaucoup, fille, voulez, jour, monsieur, regarde, soir, nom, allons, gens, mère, vite, nos, donc, cet, arrive, peur, cela, pense, air, appelle, quelle, plaît, tête, arrête, prendre, prends, maison, problème, bonjour, parle, raison, maman, ouais, savez, connais, choses, jusqu, assez, voulais, coup, aider, moment, est-ce qu, hé, partir, sait, hommes, chance, heure, elles, combien, jours, tant, hein, pu, pouvez, amour, venir, ni, comprends, travail, putain, trouver, idée, passer, vient, truc, ah, aujourd'hui, chercher, enfants, quelques, laissez, question, sang, gars, pourrait, famille, type, mec, rester, venez, regardez, eux, voici, frère, trouve, ville, longtemps, vont, histoire, yeux, heures, seulement, cas, attendez, main, importe, mourir, eh, côté, là-bas, devrais, corps, eau, mois, suite, pourrais, savais, genre, chaque, prie, excusez, mettre, aimerais, loin, matin, espère, chef, train, capitaine, écoutez, donner, jouer, prenez, endroit, enfant, devrait, femmes, façon, enfin, arrêtez, celui, film, coeur, guerre, plutôt, bientôt, hier, croire, demandé, souviens, arrêter, mari, pensais, mains, dirait, docteur, tomber, pays, confiance, ainsi, voyez, essaie, vérité, école, papa, d'abord, font, numéro, cul, fera, arriver, filles, ceux, suffit, voudrais, bébé, croyais, prend, appeler, penses, tellement, mot, sors, voulait, presque, attendre, semaine, années, tour, route, ira, journée, jeu, bureau, plaisir, ceci, garçon, croyez, devez, service, rentrer, autant, reviens, entrer, aimes, madame, attend, voit, état, pensez, verre, dollars, dès, adore, parce qu, ferai, entends, chien, facile, soeur, mariage, seigneur, cinq, parles, vaut, cherche, retour, avis, an, payer, bras, lieu, sécurité, rentre, messieurs, pensé, commence, parfois, esprit, exactement, partout, patron, montrer, parlez, génial, veulent, colonel, essayer, choix, travaille, perdre, ordre, faisait, travailler, occupe, prison, tôt, changer, l'un, sûrement, scène, entendre, rendre, regarder, donnez, retard, difficile, restez, ferais, penser, ouvre, peuvent, bonsoir, connaissez, allait, bizarre, fric, tirer, semble, dormir, six, sauver, espèce, problèmes, agent, rappelle, sûre, trucs, année, devait, celle, prix, questions, essayé, ensuite, visage, battre, face, pouvais, souvent, avion, âge, carte, rapport, oncle, agit, salle, faim, comprendre, connaît, garder, appelez, paix, chemin, voyons, photo, oeil, hôpital, dix, erreur, hôtel, fallait, message, allé, honneur, bordel, cheveux, voix, bateau, ciel, professeur, allais, mission, acheter, début, propos, pièce, ailleurs, tenez, sac, flics, télé, revenir, libre, situation, gagner, lumière, retrouver, pieds, sortez, effet, apprendre, jolie, ordres, monte

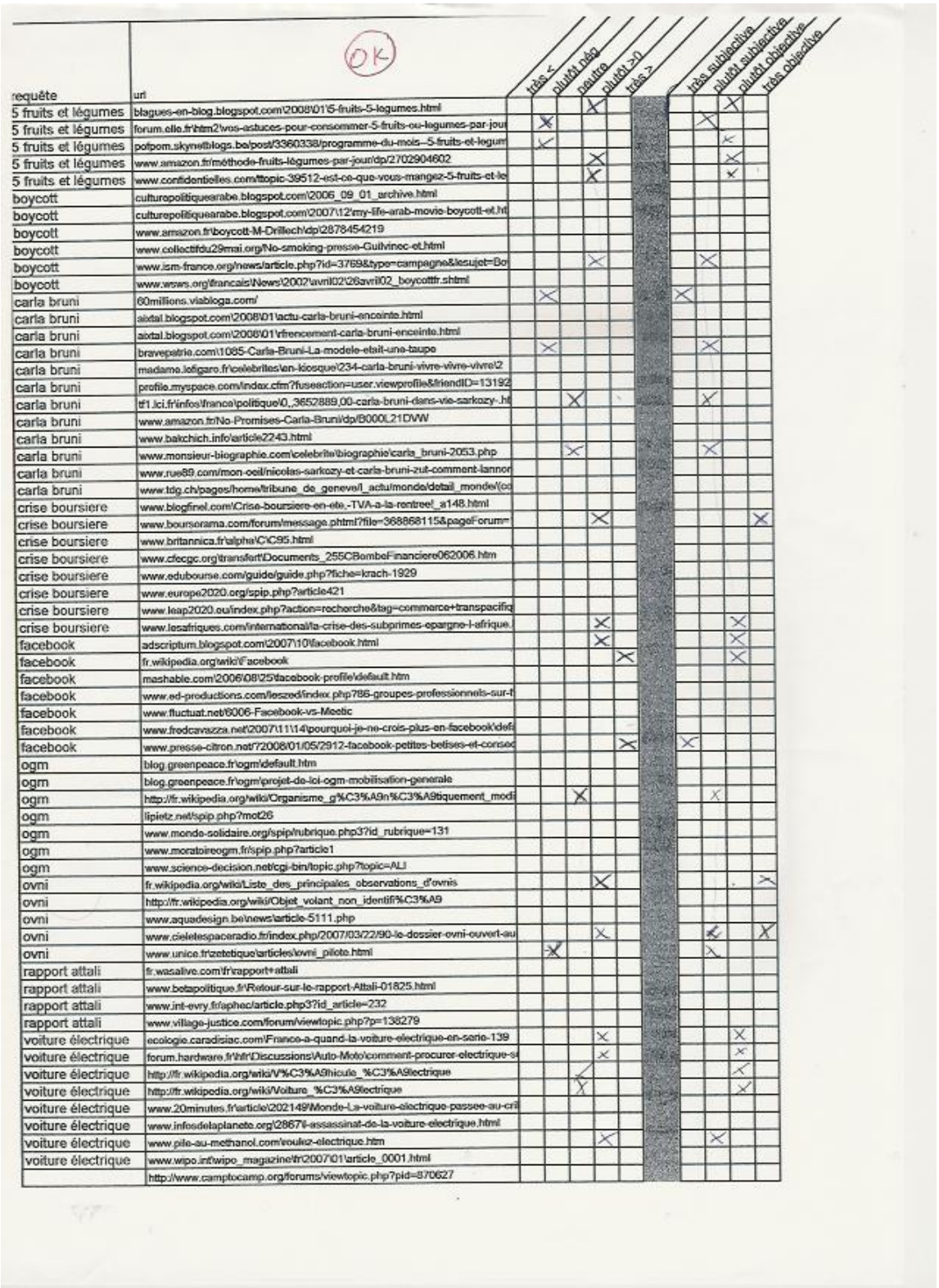
Liste des mots vides classés par ordre alphabétique et présents dans le corpus de film

à, accord, acheter, adore, âge, agent, agit, ah, aider, ailleurs, aime, aimerais, aimes, ainsi, air, allais, allait, allé, allons, alors, amour, an, année, années, ans, appeler, appelez, appelle, apprendre, arrête, arrêter, arrêtez, arrive, arriver,

assez, attend, attendez, attendre, au, aujourd'hui, aussi, autant, aux, avion, avis, bateau, battre, beaucoup, bébé, besoin, bientôt, bizarre, bonjour, bonsoir, bordel, bras, bureau, capitaine, carte, cas, ceci, cela, celle, celui, cet, cette, ceux, chance, changer, chaque, chef, chemin, cherche, chercher, cheveux, chez, chien, choix, chose, choses, ciel, cinq, coeur, colonel, combien, commence, comprendre, comprends, confiance, connais, connaissez, connaît, corps, côté, coup, croire, crois, croyais, croyez, cul, d'abord, dans, de, début, demandé, des, dès, deux, devait, devez, devrais, devrait, dieu, difficile, dirait, dis, dix, docteur, dois, dollars, donc, donner, donnez, dormir, du, eau, école, écoutez, effet, eh, elles, encore, endroit, enfant, enfants, enfin, ensuite, entendre, entends, entrer, erreur, espèce, espère, esprit, essaie, essayé, essayer, est-ce qu, est-ce que, et, état, eux, exactement, excusez, face, facile, façon, faim, faire, fais, faisait, fallait, famille, faut, femme, femmes, fera, ferai, ferais, fille, filles, film, flics, fois, font, frère, fric, gagner, garçon, garder, gars, génial, genre, gens, guerre, hé, hein, heure, heures, hier, histoire, homme, hommes, honneur, hôpital, hôtel, idée, ils, importe, ira, jamais, je, jeu, jolie, jouer, jour, journée, jours, jusqu, là-bas, laissez, libre, lieu, loin, longtemps, lumière, l'un, madame, main, mains, maison, maman, mari, mariage, matin, me, mec, mère, mes, message, messieurs, mettre, mission, mois, moment, mon, monde, monsieur, monte, montrer, mot, mourir, ni, nom, nos, nous, numéro, occupe, oeil, oh, on, oncle, ordre, ordres, ou, ouais, oui, ouvre, paix, papa, parce qu, parfois, parle, parles, parlez, partir, partout, passer, patron, payer, pays, pensais, pense, pensé, penser, penses, pensez, perdre, père, peur, peut, peut-être, peuvent, peux, photo, pièce, pieds, plaisir, plaît, plutôt, pourrais, pourrait, pouvais, pouvez, prend, prendre, prends, prenez, presque, prie, prison, prix, problème, problèmes, professeur, propos, pu, putain, quel, quelle, quelques, quelqu'un, question, questions, raison, rappelle, rapport, regarde, regarder, regardez, rendre, rentre, rentrer, rester, restez, retard, retour, retrouver, revenir, reviens, route, sa, sac, sais, sait, salle, sang, sans, sauver, savais, savez, scène, sécurité, seigneur, semaine, semble, service, ses, seulement, situation, six, soeur, soir, sors, sortez, souvent, souviens, suffit, suite, sûr, sûre, sûrement, ta, tant, te, télé, tellement, temps, tenez, tes, tête, tirer, toi, tomber, tôt, toujours, tour, train, travail, travaille, travailler, très, trop, trouve, trouver, truc, trucs, type, va, vais, vas, vaut, venez, venir, vérité, verre, veulent, veut, veux, vie, viens, vient, ville, visage, vite, voici, voilà, voir, vois, voit, voix, vont, vos, voudrais, voulais, voulait, voulez, vous, voyez, voyons, vraiment, yeux

13. Annexe 13 : exemple de fiche remplie

Exemple de fiche remplie, dans laquelle on demandait aux juges de qualifier un corpus de 63 pages web selon le critère de valence et de niveau de subjectivité.



requête	url	très <	plutôt nég	neutre	plutôt >0	très >	très subjective	plutôt subjective	plutôt objective	très objective
5 fruits et légumes	blagues-en-blog.blogspot.com\2008\01\5-fruits-5-legumes.html			X				X		
5 fruits et légumes	forum.elle.fr\htm2\vos-astuces-pour-consommer-5-fruits-ou-legumes-par-jou	X					X			
5 fruits et légumes	potpom.skynetblogs.be/post/3360338/programme-du-mois--5-fruits-et-legum	X						X		
5 fruits et légumes	www.amazon.fr/méthode-fruits-légumes-par-jour/dp/2702904602			X				X		
5 fruits et légumes	www.confidentielles.com/ttopic-39512-est-ce-que-vous-mangez-5-fruits-et-le			X				X		
boycott	culturepolitiquearabe.blogspot.com\2006_09_01_archive.html									
boycott	culturepolitiquearabe.blogspot.com\2007\12\my-life-arab-movie-boycott-et.ht									
boycott	www.amazon.fr\boycott-M-Drillech\dp\2878454219									
boycott	www.collectifdu29mai.org/No-smoking-presse-Guilvinec-et.html									
boycott	www.ism-france.org/news/article.php?id=3769&type=campagne&lesujet=Bo			X			X			
boycott	www.wsws.org/francais/News/2002\avril02\26avril02_boycotttfr.shtml									
carla bruni	60millions.viabloga.com/	X					X			
carla bruni	aixtal.blogspot.com\2008\01\actu-carla-bruni-enceinte.html									
carla bruni	aixtal.blogspot.com\2008\01\rfrencement-carla-bruni-enceinte.html									
carla bruni	bravepatrie.com\1085-Carla-Bruni-La-modele-etait-une-taupe	X					X			
carla bruni	madame.lefigaro.fr\celebrites\en-kiosque\234-carla-bruni-vivre-vivre-vivre\2									
carla bruni	profile.myspace.com/index.cfm?fuseaction=user.viewprofile&friendID=13192									
carla bruni	tf1.lci.fr/infos/france/politique/0,,3652889,00-carla-bruni-dans-vie-sarkozy-.ht		X				X			
carla bruni	www.amazon.fr/No-Promises-Carla-Bruni/dp/B000L21DVW									
carla bruni	www.baikchich.info/article2243.html									
carla bruni	www.monsieur-biographie.com\celebrite\biographie\carla_bruni-2053.php		X				X			
carla bruni	www.rue89.com/mon-oeil/nicolas-sarkozy-et-carla-bruni-zut-comment-lannor									
carla bruni	www.tdg.ch/pages/home/tribune_de_geneve/l_actu/monde/detail_monde/(co									
crise boursiere	www.blogfinel.com/Crise-boursiere-en-ete,-TVA-a-la-rentree!_a148.html									
crise boursiere	www.boursorama.com/forum/message.phtml?file=368868115&pageForum=			X						X
crise boursiere	www.britannica.fr/alpha/C%C9S.html									
crise boursiere	www.cfecgc.org/transfert/Documents_255CBombeFinanciere062006.htm									
crise boursiere	www.edubourse.com/guide/guide.php?fiche=krach-1929									
crise boursiere	www.europe2020.org/spip.php?article421									
crise boursiere	www.leap2020.eu/index.php?action=recherche&tag=commerce+transpacifiq									
crise boursiere	www.lesafriques.com/international/la-crise-des-subprimes-epargne-l-afrique.			X				X		
facebook	adscriptum.blogspot.com\2007\10\facebook.html			X				X		
facebook	fr.wikipedia.org/wiki/Facebook				X			X		
facebook	mashable.com\2006\08\25\facebook-profile\default.htm									
facebook	www.ed-productions.com/leszed/index.php?86-groupes-professionnels-sur-f									
facebook	www.fluctuat.net/6006-Facebook-vs-Meetic									
facebook	www.fredcavazza.net\2007\11\14\pourquoi-je-ne-crois-plus-en-facebook\defi									
facebook	www.presse-citron.net/?2008/01/05/2912-facebook-petites-betises-et-consec				X		X			
ogm	blog.greenpeace.fr\ogm\default.htm									
ogm	blog.greenpeace.fr\ogm\projet-de-loi-ogm-mobilisation-generale									
ogm	http://fr.wikipedia.org/wiki/Organisme_g%C3%A9n%C3%A9tiquement_modi		X					X		
ogm	lipietz.net/spip.php?mot26									
ogm	www.monde-solidaire.org/spip/rubrique.php3?id_rubrique=131									
ogm	www.moratoireogm.fr/spip.php?article1									
ogm	www.science-decision.net/cgi-bin/topic.php?topic=ALI									
ovni	fr.wikipedia.org/wiki/Liste_des_principales_observations_d'ovnis			X						X
ovni	http://fr.wikipedia.org/wiki/Objet_volant_non_identifi%C3%A9									
ovni	www.aquadesign.be\news\article-5111.php									
ovni	www.cieletespaceradio.fr/index.php/2007/03/22/90-le-dossier-ovni-ouvert-au			X				X		X
ovni	www.unice.fr/zetetique/articles/ovni_pilote.html		X					X		
rapport attali	fr.wasalive.com/fr/rapport+attali									
rapport attali	www.betapolitique.fr\Retour-sur-le-rapport-Attali-01825.html									
rapport attali	www.int-evry.fr/aphec/article.php3?id_article=232									
rapport attali	www.village-justice.com/forum/viewtopic.php?p=138279									
voiture électrique	ecologie.caradisiac.com\France-a-quand-la-voiture-electrique-en-serie-139			X				X		
voiture électrique	forum.hardware.fr\hfr\Discussions\Auto-Moto\comment-procurer-electrique-s			X				X		
voiture électrique	http://fr.wikipedia.org/wiki/V%C3%A9hicule_%C3%A9lectrique							X		
voiture électrique	http://fr.wikipedia.org/wiki/Voiture_%C3%A9lectrique			X				X		
voiture électrique	www.20minutes.fr\article\202149\Monde-La-voiture-electrique-passee-au-cri									
voiture électrique	www.infosdelaplanete.org\2867\l-assassinat-de-la-voiture-electrique.html									
voiture électrique	www.pile-au-methanol.com/voulez-electrique.htm									
voiture électrique	www.wipo.int\wipo_magazine\fr\2007\01\article_0001.html									
	http://www.camptocamp.org/forums/viewtopic.php?pid=870627									

14. Annexe 14 : Indicateur d'accessibilité de 21 sites brésiliens appartenant au répertoire national des sites accessibles

	A		AA		AAA					
	ERR DIFF	TOTAL ERR	ERR DIFF	TOTAL ERR	ERR DIFF	TOTAL ERR	note A	note AA	note AAA	note totale
http://www.bancazeebia.com.br/index.htm	0	0	1	1	0	0	30	24	30	19,00
http://www.brazilprime.com	0	0	0	0	1	1	30	30	24	19,40
http://www.anaemauro.arg.br	1	2	3	4	1	1	23	11	24	13,43
http://www.planetaeducacao.com.br	0	0	1	1	1	1	30	24	24	18,40
http://www.dse.cepel.br	1	1	6	15	1	1	24	0	24	12,00
http://animeru.webng.com	1	1	0	0	1	1	24	30	24	17,00
http://www.presidencia.gov.br	0	0	4	39	2	14	30	0	6	12,60
http://www.fapespa.pa.gov.br	0	0	3	15	0	0	30	0	30	15,00
http://www.tekroll.com.br/	0	0	2	8	1	4	30	12	21	16,10
http://www.conexaotocantins.com.br	0	0	0	0	1	1	30	30	24	19,40
http://www.ilhainternet.com.br/	0	0	0	0	1	1	30	30	24	19,40
http://www.adiprofar.com.br	0	0	0	0	1	1	30	30	24	19,40
http://www.acuity.com.br/	1	1	6	30	2	11	24	0	9	10,50
http://www.deficienteonline.com.br/principal/I	0	0	6	93	2	37	30	0	0	12,00
http://www.edificiobrasil.com.br	0	0	0	0	0	0	30	30	30	20,00
http://www.sincab.org	0	0	3	34	2	17	30	0	3	12,30
http://www.capoeirabrasilamazonia.com	0	0	0	0	1	1	30	30	24	19,40
http://www.kinghost.com.br	0	0	3	5	1	1	30	10	24	16,07
http://www.soperfume.com.br	0	0	1	7	1	1	30	18	24	17,40
http://www.transcartransportes.com.br/index.	0	0	2	8	1	1	30	12	24	16,40
http://www.clickaju.com/home.php	0	0	0	0	0	0	30	30	30	20,00
score moyen										16,44

15. Annexe 15 : comparaison de l'indicateur d'accessibilité de sites répondant au label accessiweb et de sites d'entreprises du Cac 40

		A		AA		AAA						
		ERR DIFF	TOTAL ERR	ERR DIFF	TOTAL ERR	ERR DIFF	TOTAL ERR	label accessiweb				
	ENTREPRISES LABELLISEES ACCESSIWEB								note A	note AA	note AAA	note totale
1	www.tcl.fr	0	0	2	13	0	0	Argent	30	7	30	16,17
2	www.nouvelleuniversite.gouv.fr	0	0	1	5	0	0	or	30	20	30	18,33
3	www.micropole-univers.com	0	0	1	2	0	0	bronze	30	23	30	18,83
4	www.integrance.fr	0	0	2	2	1	1	or	30	18	24	17,40
5	www.ter-sncf.com	0	0	1	1	1	1	Argent	30	24	24	18,40
6	www.halpades.fr	0	0	4	14	0	0	bronze	30	0	30	15,00
7	www.semitag	0	0	2	4	0	0	or	30	16	30	17,67
8	www.meylan-	0	0	3	6	0	0	or	30	9	30	16,50
9	www.gazdefra	0	0	6	61	1	4	bronze	30	0	21	14,10
10	www.business	0	0	2	6	0	0	or	30	14	30	17,33
11	www.handicar	0	0	3	26	1	12	bronze	30	0	13	13,30
12	www.adapei70.org	0	0	1	1	0	0	or	30	24	30	19,00
13	www.equallia.com	0	0	2	4	0	0	or	30	16	30	17,67
14	http://www.handicap13.fr/handicap13/CG13/	0	0	3	11	1	7	bronze	30	4	18	14,47
15	www.laposte.fr/courrierinternational/	0	0	5	20	0	0	argent	30	0	30	15,00
16	www.msa.fr	0	0	3	28	1	12	bronze	30	0	13	13,30
17	www.planetarium-galilee.com	0	0	1	1	1	1	bronze	30	24	24	18,40
18	http://www.outre-mer.gouv.fr/outremer/front	0	0	2	8	1	3	bronze	30	12	22	16,20
19	www.cr-languedocroussillon.fr	0	0	3	26	0	0	Or	30	0	30	15,00
20	www.legrandchalon.fr	0	0	2	13	0	0	or	30	7	30	16,17
21	www.integrance.fr	0	0	2	2	1	1	or	30	18	24	17,40
	ENTREPRISES DU CAC 40											
1	www.ppr.com/	0	0	4	13	1	5		30	0	20	14,00
2	http://www.vivendi.com/corp/fr/home/index.php	1	11	5	14	2	9		14	0	11	6,70
3	www.credit-agricole.fr/	1	1	2	30	0	0		24	0	30	12,60
4	www.casino.fr/	0	0	2	6	1	1		30	14	24	16,73
5	http://www.agence.francetelecom.com/	0	0	3	11	0	0		30	4	30	15,67
6	http://www.michelin.fr/fr/groupe_france/home.jsp	1	1	7	84	1	12		24	0	13	10,90
7	www.tf1.fr/	1	107	12	591	1	109		0	0	0	0,00
8	www.societegenerale.fr/	0	0	4	16	1	1		30	0	24	14,40
9	www.bnpparibas.com/	1	7	7	15	1	3		18	0	22	9,40
10	www.lagardere.com/	0	0	1	1	1	1		30	24	24	18,40
11	www.carrefour.fr/	6	9	1	64	1	1		0	0	24	2,40
12	http://www.bouygues.com/fr/index.asp	1	26	6	35	2	14		0	0	6	0,60
13	http://www.accor.fr/fr/index.asp	0	0	5	38	2	18		30	0	2	12,20
14	www.lvmh.fr/	2	2	2	3	1	1		18	17	24	12,43
15	http://www.lorealfr/_fr/_fr/index.aspx	1	4	7	45	2	9		21	0	11	9,50
16	www.psa.fr/	0	0	0	0	1	1		30	30	24	19,40
17	www.axa.fr/	1	1	10	83	2	25		24	0	0	9,60
18	www.danone.com/	1	15	2	6	0	0		10	14	30	9,33
19	http://www.renault.com/renault_com/fr/main/index.aspx	1	1	2	97	0	0		24	0	30	12,60
20	www.total.com/	1	19	9	95	2	21		6	0	0	2,40
21	www.lafarge.fr/	0	0	2	10	1	1		30	10	24	16,07
22	www.agf.fr/	0	0	5	23	0	0		30	0	30	15,00
23	www.airliquide.fr/	0	0	1	1	1	1		30	24	24	18,40
24	http://www.alcatel-lucent.com/wps/portal?lu_lang_code=fr_WW	0	0	5	61	2	28		30	0	0	12,00
25	http://www.arcelormittal.com/	1	1	7	21	2	6		24	0	14	11,00
26	www.fr.capgemini.com/	0	0	4	4	1	1		30	6	24	15,40
27	www.dexia.com/	1	1	3	8	2	4		24	7	16	12,37
28	www.essilor.fr/	1	37	4	21	1	1		0	0	24	2,40
29	http://www.eads.com/1024/en/Trailer_EADS.html	1	7	5	15	1	5		18	0	20	9,20
30	www.pernod-ricard.com/	1	34	8	23	2	4		0	0	16	1,60

16. Annexe 16 : Complexité et temps de calcul

Dans ce paragraphe, nous allons nous intéresser à la question de la complexité des algorithmes mis en œuvre pour calculer les facettes et du temps de calcul effectif. Nous nous intéresserons ensuite à ces notions dans le cadre de l'intégration de ces algorithmes au sein de moteurs de recherche.
Nous allons nous attacher aux facettes de valence, subjectivité et lisibilité qui reposent sur des logiques de traitement de l'information mises en œuvre dans le cadre de ce travail. Dans le cas des autres facettes (accessibilité et centralité), nous sollicitons des routines externes qui effectuent les calculs à notre place.

a) En théorie : calcul de la complexité temporelle des algorithmes

Considérons par exemple la facette valence. Rappelons les étapes de l'algorithme :

1. Extraction des m termes présents dans la page web

La complexité de calcul de cette étape est linéaire c'est à dire en O(m).

2. Recherche de la valence de chaque terme dans le dictionnaire contenant n termes

Rechercher un terme dans un dictionnaire indexé de n termes a une complexité logarithmique c'est-à-dire en O(log n). Ainsi, pour m termes, la complexité est en O(m.log(n)).

3. Sélection de termes de valence extrême parmi m termes

Cette étape est réalisée en temps linéaire : O(m). Elle consiste en effet à retenir parmi les m termes de la page ceux réalisant la condition.

4. Calcul de moyenne pour m termes d'une page web

Ce calcul se fait en temps linéaire soit en O(m).

Ainsi, globalement, la complexité temporelle de l'algorithme est log linéaire c'est-à-dire en O(m.log(n)). Cette complexité est tout à fait « raisonnable » et

ne doit pas interdire le traitement de gros corpus de données ni de gros dictionnaires.

A cette complexité temporelle s'ajoute la complexité spatiale liée à l'occupation mémoire. Pour optimiser cette complexité, il est important de libérer dès que possible la place mémoire occupée par les données traitées. Par exemple, il est inutile de conserver en mémoire la totalité des pages du corpus : une fois une page traitée, le tableau de termes associés peut être supprimé (ainsi les pages peuvent être traitées indépendamment).

b) En pratique : temps de calcul des algorithmes

Les résultats présentés dans cette partie ont été obtenus sur une configuration matérielle de type Pentium (R) 4 CPU 3Ghz, 1,00 Go de RAM.

Nous proposons de traiter un corpus de 279 pages web comportant 591 760 mots et un dictionnaire de 35 455 mots. Reprenons les 4 étapes décrites précédemment :

1. Extraction des m termes présents dans la page web

Cette étape est réalisée avec les logiciels WebPipe Pro et TextPipe Pro
Le temps de calcul est proportionnel au nombre de mots m de la page.
L'algorithme traite en moyenne 65000 mots (la taille moyenne d'une page est en moyenne de 2000 mots) en 1 seconde.

2. Recherche de la valence des m termes dans le dictionnaire contenant n termes

Cette étape a été réalisée à l'aide du logiciel Access : à chaque page est associée une table contenant la liste des mots clés qu'elle contient (avec redondances éventuelles). Une autre table contient pour chaque mot clé (choisi comme index) sa valence. Une requête SELECT est ensuite créée pour associer à chaque mot clé de la page sa valence. Le temps de calcul de cette opération est estimé à une fraction de seconde pour la totalité des 279 pages.

3. Sélection de termes de valence extrême parmi m termes

En modifiant l'opération SELECT précédente, on peut rajouter un critère : valence > borne1 OU valence < borne2, avec borne1 et borne2 : les limites en dehors desquelles on retiendra la valence du terme pour le calcul de la moyenne. Cette opération ne modifie pas sensiblement le temps de calcul de l'étape précédente (une fraction de seconde).

4. Calcul de moyenne pour m termes d'une page web

Bien que la complexité de cette étape soit linéaire, son temps de calcul est plus élevé que celui des 2 étapes précédentes. Sur notre jeu de données, on compte prêt de 6 secondes pour calculer la valence moyenne des 279 pages. Ce qui reste tout à fait raisonnable si on se ramène au temps de calcul unitaire (2/100 de seconde par page).

c) Vers une mise en production

La démarche suivie n'est actuellement pas complètement automatisée : elle fait appel à divers outils : WebPipe pro, TextPipe pro et Access et nécessite l'intervention de l'expérimentateur qui doit ouvrir tour à tour les divers fichiers générés.
L'intégration de ces différentes étapes (du parsing au traitement des données) dans un même outil permettrait d'automatiser entièrement le calcul des différents indices.

Dans le prototype actuel, le choix a été porté par simplicité sur l'outil Access mais des solutions plus orientées web pourraient être avantageusement utilisées en production.

Pour conclure, l'algorithme proposé a une faible complexité (log linéaire), ce qui se traduit par un temps de calcul très raisonnable même lorsque la taille de la page ou celle du dictionnaire augmente.

Ceci laisse espérer que l'algorithme puisse être calculé en temps réel (c'est-à-dire à la volée) pour chaque page avec un temps de traitement de l'ordre de la fraction de seconde.

Pour une intégration des divers indices au sein d'un moteur de recherche, ce calcul doit être réalisé off line (à l'occasion d'une Google Dance !). Aussi, il n'induit pas un gros surcoût de temps de calcul, d'autant que les pages sont déjà parsées par les moteurs.